

# JOURNAL OF APPLIED PSYCHOLOGY

EDWIN A. FLEISHMAN, EDITOR  
*American Institutes for Research*

## CONSULTING EDITORS

- EARL A. ALLUISI  
*University of Louisville*
- BERNARD M. BASS  
*University of Rochester*
- DONALD E. BROADBENT  
*Applied Psychology Unit, Medical Research Council, Cambridge, England*
- JOHN P. CAMPBELL  
*University of Minnesota*
- PIETER J. D. DRENTH  
*Free University, Amsterdam, The Netherlands*
- MARVIN D. DUNNETTE  
*University of Minnesota*
- FRANK FRIEDLANDER  
*Case Western Reserve University*
- ALBERT S. GLICKMAN  
*American Institutes for Research, Washington, D. C.*
- LEONARD V. GORDON  
*State University of New York at Albany*
- L. RICHARD HOFFMAN  
*University of Chicago*
- PAUL HORST  
*University of Washington*
- EDWIN A. LOCKE  
*University of Maryland*
- JOSEPH E. McGRATH  
*University of Illinois, Urbana*
- ROBERT B. MILLER  
*IBM, Poughkeepsie, New York*
- WILLIAM G. MOLLENKOPF  
*Procter & Gamble Company, Cincinnati, Ohio*
- WILLIAM A. OWENS  
*University of Georgia*
- ROBERT PERLOFF  
*University of Pittsburgh*
- FRANCIS J. PILGRIM  
*Chicago, Illinois*
- LYMAN W. PORTER  
*University of California, Irvine*
- ANNE ROE  
*University of Arizona*
- STANLEY E. SEASHORE  
*University of Michigan*
- MARY L. TENOPYR  
*AT&T, New York*
- HARRY TRIANDIS  
*University of Illinois, Urbana*
- DON TRUMBO  
*Pennsylvania State University*
- J. E. UHLANER  
*U. S. Army Research Institute for the Behavioral and Social Sciences, Arlington, Virginia*
- ALEXANDER G. WESMAN  
*The Psychological Corporation, New York*

VOLUME 57, 1973

HAROLD P. VAN COTT  
Managing Editor

ANITA DE VIVO  
Executive Editor and  
General Manager

BARBARA HOBBS  
Supervising  
Technical Editor

ANN G. CORBETT  
Technical Editor

HENRY S. SALMON  
Advertising and  
Business Manager

ANNE REDMAN  
Subscription  
Manager

PUBLISHED BIMONTLY BY  
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.  
1200 Seventeenth St., N. W., Washington, D. C. 20036

Copyright © 1973 by the American Psychological Association, Inc.

6-3-81

# JOURNAL OF APPLIED PSYCHOLOGY

## ACKNOWLEDGMENTS

In addition to the Editorial Board of Consulting Editors, the following individuals assisted in the review of manuscripts submitted to the JOURNAL from January 1972 through June 1972. The Editor gratefully acknowledges their contributions.

- |                       |                             |
|-----------------------|-----------------------------|
| HERBERT I. ABELSON    | JACOB JACOBY                |
| DOROTHY C. ADKINS     | EDMUND T. KLEMMER           |
| ALEXANDER W. ASTIN    | SAMUEL S. KOMORITA          |
| MELANY E. BAEHR       | JANOS B. KOPLYAY            |
| ALAN R. BASS          | ABRAHAM K. KORMAN           |
| HAROLD P. BECHTOLDT   | FRANK J. LANDY              |
| MILTON R. BLOOD       | GARDNER LINDZEV             |
| DARRELL R. BOCK       | WILLIAM A. MCCLELLAND       |
| TIMOTHY C. BROCK      | WILBERT McKEACHIE           |
| HUBERT E. BROGDEN     | HERBERT H. MEYER            |
| DONALD T. CAMPBELL    | ANGELO MIRABELLA            |
| ROBERT E. CARLSON     | DONALD C. PELZ              |
| RAYMOND E. CHRISTAL   | EVELYN PERLOFF              |
| NORMAN CLIFF          | HJALMAR ROSEN               |
| ANDREW L. COMREY      | MELVIN H. RUDOV             |
| ROYER F. COOK         | JEROME M. SATTLER           |
| TIMOTHY W. COSTELLO   | EDGAR H. SCHEIN             |
| JOHN O. CRITES        | BENJAMIN SCHNEIDER          |
| LEE J. CRONBACH       | RALPH M. STOGDILL           |
| H. PETER DACHLER      | DONALD E. SUPER             |
| SEYMOUR FISHER        | ABRAHAM TESSER              |
| THEODORE W. FORBES    | GEORGE C. THEOLOGUS         |
| MONROE FRIEDMAN       | ROBERT L. THORNDIKE         |
| BASIL S. GEORGOPoulos | DAVID V. TIEDEMAN           |
| GOLDINE C. GLESER     | STEVEN G. VANDENBERG        |
| PAUL S. GOODMAN       | VICTOR H. VROOM             |
| J. P. GUILFORD        | JOE H. WARD, JR.            |
| ROBERT M. GUION       | JOSEPH WEITZ                |
| RICHARD J. HACKMAN    | KENNETH N. WEXLEY           |
| CLIFFORD P. HAHN      | GEORGE R. WHEATON           |
| MILTON D. HAKEL       | HERMAN A. WITKIN            |
| EUGENE HEATON, JR.    | LAWRENCE S. WRIGHTSMAN, JR. |
| HARRY HELSON          | GARY A. YUKL                |
| JOHN L. HOLLAND       | ALBERT ZAVALA               |
| EDWIN P. HOLLANDER    | SHELDON ZEDECK              |

## CONTENTS OF VOLUME 57

Aiden, David G. <i>See</i> Wedell, Jacelyn.	
Arvey, Richard D. <i>See</i> Campbell, John P.	
Azen, Stanley P., Snibbe, Homa M., and Montgomery, Hugh R. A longitudinal predictive study of success and performance of law enforcement officers.....	190
Bakeman, Roger. <i>See</i> Helmreich, Robert.	
Baskett, Glen D. Interview decisions as determined by competency and attitude similarity.....	343
Bass, Alan R., and Turner, John N. Ethnic group differences in relationships among criteria of job performance.....	101
Bigelow, Douglas A., and Driscoll, Richard H. Effect of minimizing coercion on the rehabilitation of prisoners.....	10
Biglan, Anthony. Relationships between subject matter characteristics and the structure and output of university departments.....	204
Biglan, Anthony. The characteristics of subject matter in different academic areas.....	195
Blout, Harry D. <i>See</i> Cammalleri, Joseph A.	
Brislin, Richard W. <i>See</i> Sinaiko, H. Wallace.	
Burns, W. J. <i>See</i> Jones, Richard R.	
Butler, Richard P. Effects of signed and unsigned questionnaires for both sensitive and nonsensitive items.....	348
Cahoon, Richard L. Auditory vigilance under hypoxia.....	350
Cameron, Paul, and Robertson, Donald. Effect of home environment tobacco smoke on family health.....	142
Cammalleri, Joseph A., Hendrick, Hal W., Pittman, Wayne C., Jr., Blout, Harry D., and Prather, Dirk C. Effects of different leadership styles on group accuracy.....	32
Campbell, John P., Dunnette, Marvin D., Arvey, Richard D., and Hellervik, Lowell V. The development and evaluation of behaviorally based rating scales.....	15
Carlsmith, J. Merrill. <i>See</i> Doob, Anthony N.	
Chemers, Martin M. <i>See</i> Rice, Robert W.	
Claudy, John G. <i>See</i> Richards, James M., Jr.	
Copeland, Charles. <i>See</i> Kipnis, David.	
Cravens, David W., and Woodruff, Robert B. An approach for determining criteria of sales performance.....	242
De Leo, Philip J. <i>See</i> Pritchard, Robert D.	
Donnelly, James H., Jr., Etzel, Michael J., and Roeth, Scott. The relationship between consumers' category width and trial of new products.....	335
Doob, Anthony N., Freedman, Jonathan L., and Carlsmith, J. Merrill. Effects of sponsor and prepayment on compliance with a mailed request.....	346
Driscoll, Richard H. <i>See</i> Bigelow, Douglas A.	
Dunnette, Marvin D. <i>See</i> Campbell, John P.	
Dunnette, Marvin D. <i>See</i> Jorgenson, Dale O.	
Etzel, Michael J. <i>See</i> Donnelly, James H., Jr.	
Farr, James L. Response requirements and primacy-recency effects in a simulated selection interview.....	228
Fenster, C. Abraham, and Locke, Bernard. Neuroticism among policemen: An examination of police personality.....	358
Fiedler, Fred E. Predicting the effects of leadership training and experience from the contingency model: A clarification.....	110
Fossum, John A. An application of techniques to shorten tests and increase validity.....	90
Fox, William M., Hill, Walter A., and Guertin, Wilson H. Dimensional analysis of the least preferred co-worker scales.....	192
Freedman, Jonathan L. <i>See</i> Doob, Anthony N.	
Gannon, Martin J., and Hendrickson, D. Hunt. Career orientation and job satisfaction among working wives.....	339
Gatewood, Robert D., and Perloff, Robert. An experimental investigation of three methods of providing weight and price information to consumers.....	81
Goldman, Roy D., Platt, Bruce B., and Kaplan, Robert B. Dimensions of attitudes toward technology.....	184
Goodale, James G. Effects of personal background and training on work values of the hard-core unemployed.....	1
Guertin, Wilson H. <i>See</i> Fox, William M.	
Guilford, Joan S. Prediction of accidents in a standardized home environment.....	306
Harlan, Anne. <i>See</i> Kerr, Steven.	
Harlow, Dorothy N. Professional employees' preference for upward mobility.....	137
Hegarty, W. Harvey. <i>See</i> Rosen, Benson.	
Hellervik, Lowell V. <i>See</i> Campbell, John P.	
Helmreich, Robert, Bakeman, Roger, and Radloff, Roland. The life history questionnaire as a predictor of performance in Navy diver training.....	148
Hendrick, Hal W. <i>See</i> Cammalleri, Joseph A.	
Hendrickson, D. Hunt. <i>See</i> Gannon, Martin J.	

Herman, Jeanne B., and Hulin, Charles L. Managerial satisfactions and organizational roles: An investigation of Porter's Need Deficiency Scales.....	118
Hill, Walter A. <i>See</i> Fox, William M.	
Hoffmann, Errol R. <i>See</i> Macdonald, Wendy A.	
Hulin, Charles L. <i>See</i> Herman, Jeanne B.	
Jerdee, Thomas H. <i>See</i> Rosen, Benson.	
Johnson, Raymond H. <i>See</i> Schmidt, Frank L.	
Jones, Richard R., and Burns, W. J. Volunteer satisfaction with in-country training for the Peace Corps: Reanalyses and extended findings.....	92
Jorgenson, Dale O., Dunnette, Marvin D., and Pritchard, Robert D. Effects of the manipulation of a performance-reward contingency on behavior in a simulated work setting.....	271
Kanungo, Rabindra N., and Pang, Sam. Effects of human models on perceived product quality.....	172
Kao, Henry S. R. The dynamic role of eye-head angular displacements in human vehicular guidance.....	320
Kaplan, Robert B. <i>See</i> Goldman, Roy D.	
Kerr, Steven, and Harlan, Anne. Predicting the effects of leadership training and experience from the contingency model: Some remaining problems.....	114
Kipnis, David, Silverman, Arnold, and Copeland, Charles. Effects of emotional arousal on the use of supervised coercion with black and union employees.....	38
Langdale, John A., and Weitz, Joseph. Estimating the influence of job information on interviewer agreement	23
Levine, Jerrold M. Information seeking with conflicting and irrelevant inputs.....	74
Locke, Bernard. <i>See</i> Fenster, C. Abraham.	
Macdonald, Wendy A., and Hoffmann, Errol R. The recognition of road pavement messages.....	314
Massey, Iris H. <i>See</i> Mullins, Cecil J.	
McGuire, Frederick L. The nature of bias in official accident and violation records.....	300
Mitchell, Terence R., and Nebeker, Delbert M. Expectancy theory predictions of academic effort and performance.....	61
Montgomery, Hugh R. <i>See</i> Azen, Stanley P.	
Mudd, Samuel. <i>See</i> Pohlman, Alan.	
Mullins, Cecil J., and Massey, Iris H. An evaluation of item-by-item test administration.....	188
Nebeker, Delbert M. <i>See</i> Mitchell, Terence R.	
Neeley, James D., Jr. A test of the need gratification theory of job satisfaction.....	86
Nystrom, Paul C. Equity theory and career pay: A computer simulation approach.....	125
O'Reilly, Charles A., III, and Roberts, Karlene H. Job satisfaction among whites and nonwhites: A cross-cultural approach.....	295
Orvik, James M. The predictive validity of premilitary performance ratings by high school personnel.....	88
Pang, Sam. <i>See</i> Kanungo, Rabindra N.	
Perloff, Robert. <i>See</i> Gatewood, Robert D.	
Pinder, Craig C. Statistical accuracy and practical utility in the use of moderator variables.....	214
Pittman, Wayne C., Jr. <i>See</i> Cammalleri, Joseph A.	
Platt, Bruce B. <i>See</i> Goldman, Roy D.	
Pohlman, Alan, and Mudd, Samuel. Market image as a function of consumer group and product type: A quantitative approach.....	167
Prather, Dirk C. Prompted mental practice as a flight simulator.....	353
Prather, Dirk C. <i>See</i> Cammalleri, Joseph A.	
Pritchard, Robert D., and De Leo, Philip J. Experimental test of the valence-instrumentality relationship in job performance.....	264
Pritchard, Robert D., and Sanders, Mark S. The influence of valence, instrumentality, and expectancy on effort and performance.....	55
Pritchard, Robert D. <i>See</i> Jorgenson, Dale O.	
Radloff, Roland. <i>See</i> Helmreich, Robert.	
Reid, Fraser. <i>See</i> Wright, Patricia.	
Rice, Robert W., and Chemers, Martin M. Predicting the emergence of leaders using Fiedler's contingency model of leadership effectiveness.....	281
Richards, James M., Jr., and Claudy, John G. Does farm practice adoption involve a general trait?.....	360
Roach, Darrell. <i>See</i> Waters, L. K.	
Roberts, Karlene H. <i>See</i> O'Reilly, Charles A., III.	
Robertson, Donald. <i>See</i> Cameron, Paul.	
Roeth, Scott. <i>See</i> Donnelly, James H., Jr.	
Rosen, Benson, and Jerdee, Thomas H. The influence of sex-role stereotypes on evaluations of male and female supervisory behavior.....	44
Rosen, Benson, Jerdee, Thomas H., and Hegarty, W. Harvey. Effects of participation in a simulated society on attitudes of business students.....	355
Rubinsky, Stanley, and Smith, Nelson. Safety training by accident simulation.....	68
Runyon, Kenneth E. Some interactions between personality variables and management styles.....	288

Sanders, Mark S. <i>See</i> Pritchard, Robert D.	
Sanders, Raymond E. <i>See</i> Wexley, Kenneth N.	
Sands, William A. A method for evaluating alternative recruiting-selection strategies: The CAPER model	222
Schein, Virginia Ellen. The relationship between sex role stereotypes and requisite management characteristics	95
Schmidt, Frank L., and Johnson, Raymond H. Effect of race on peer ratings in an industrial situation....	237
Schneider, Benjamin. The perception of organizational climate: The customer's view.....	248
Seiler, Dale A. <i>See</i> Williams, William E.	
Sherman, John. <i>See</i> Weiss, Howard.	
Shiflett, Samuel C. Performance effectiveness and efficiency under different dyadic work strategies.....	257
Silverman, Arnold. <i>See</i> Kipnis, David.	
Sinaiko, H. Wallace, and Brislin, Richard W. Evaluating language translations: Experiments on three assessment methods.....	328
Smith, Nelson. <i>See</i> Rubinsky, Stanley.	
Smith, Steve. <i>See</i> Worthing, Parker M.	
Snibbe, Homa M. <i>See</i> Azen, Stanley P.	
Tscheulin D. Leader behavior measurement in German industry.....	28
Turner, John N. <i>See</i> Bass, Alan R.	
Venkatesan, M. <i>See</i> Worthing, Parker M.	
Waters, L. K., and Roach, Darrell. Job attitudes as predictors of termination and absenteeism: Consistency over time and across organizational units.....	341
Wedell, Jacelyn, and Alden, David G. Color versus numeric coding in a keeping-track task: Performance under varying load conditions.....	154
Weiss, Howard, and Sherman, John. Internal-external control as a predictor of task effort and satisfaction subsequent to failure.....	132
Weitz, Joseph. <i>See</i> Langdale, John A.	
Wexley, Kenneth N., Sanders, Raymond E., and Yukl, Gary A. Training interviewers to eliminate contrast effects in employment interviews.....	233
Williams, William E., and Seiler, Dale A. Relationship between measures of effort and job performance	49
Woodruff, Robert B. <i>See</i> Cravens, David W.	
Worthing, Parker M., Venkatesan, M., and Smith, Steve. Personality and product use revisited: An exploration with the personality research form.....	179
Wright, Patricia, and Reid, Fraser. Written information: Some alternatives to prose for expressing the outcomes of complex contingencies.....	160
Yukl, Gary A. <i>See</i> Wexley, Kenneth N.	

## INFORMATION FOR CONTRIBUTORS

*Style.* A professional article should possess certain characteristics: (a) conciseness and an apparent respect for reader time; (b) unambiguous and simple vocabulary with technical and erudite words used only when simpler ones would obviously be inadequate; (c) conformity to accepted technical style in tables, terminology, and references; (d) conclusions that are clearly related to the evidence presented. The reader should be led, step by step, from a statement of problem or purpose, through analysis of evidence, to conclusions and implications. Authors are encouraged to consult the little book entitled *The Elements of Style* by W. Strunk, Jr., and E. B. White (New York, Macmillan, 1962).

*Criteria for evaluation.* Manuscripts will be evaluated on the basis of several criteria, including: (a) significance in contributing new knowledge to the field, (b) technical adequacy, (c) appropriateness for the *Journal of Applied Psychology*, and (d) clarity of presentation.

*Format.* Manuscripts must be prepared in the format described in the *Publication Manual of the American Psychological Association* (1967 Revision), obtainable for \$1.50 from the American Psychological Association. Articles not prepared in this manner cannot be reviewed. Special attention should be given to the section on typing the manuscript (p. 48) and to the sections on tables, figures, and references. Note that *all copy must be double spaced*, including references, title, figure captions, etc. The senior author's last name should appear on each page in the upper left-hand corner except when blind review is required. In reference lists, give journal titles in full; do not abbreviate.

*Optional blind review.* Blind reviewing may be obtained if specifically requested at the time a manuscript is first submitted. In such cases, the author's name and affiliation should appear only on a separate title page, which must be included with each copy of the manuscript. Footnotes containing information pertaining to the identity of the author or his affiliation should be on separate pages.

*Copies.* All manuscripts must be submitted in triplicate. One copy should be the original typed copy, the other two, clear carbon copies or photo reproductions, only. Authors should check the final typing carefully and retain a copy of the manuscript as a precaution against loss in the mail.

*Length.* Before preparing a manuscript, the author should check several recent issues to get an idea of the approximate length of regular articles published in the *Journal of Applied Psychology*. (One printed page equals roughly three double-spaced typewritten manuscript pages.) A few longer articles of special significance may be printed from time to time as monographs. Occasionally the *Journal* will have a section of "Short Notes" featuring brief reports on studies which make some methodological contribution or constitute an important replication.

*Abstracts.* Each copy of the manuscript must be accompanied by an abstract of 100-120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

*Figures.* The original drawing of a figure, or an 8 X 10-inch glossy print of the drawing, is needed for publication and must be submitted with the original typed copy of the manuscript. Duplicate copies of the figure may be photographic or pencil-drawn copies. Figures must be hand-lettered; typewritten lettering is not acceptable.

*Reprints.* No gratis reprints are supplied. Reprints may be ordered from the printer prior to publication.

*Supplementary material.* Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Authors should keep supporting and raw data for at least five years after publication of their article, and if applicable, offer them to the reader upon request.

# JOURNAL OF APPLIED PSYCHOLOGY

February 1973

Vol. 57, No. 1

Copyright © 1973 by the American Psychological Association, Inc.

## ARTICLES

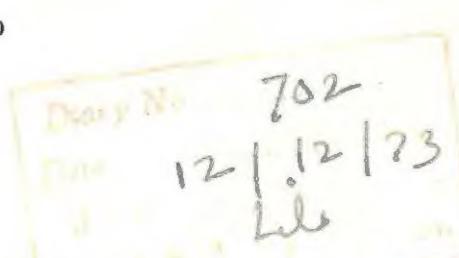
- Effects of Personal Background and Training on Work Values of the Hard-Core Unemployed *James G. Goodale* 1
- Effect of Minimizing Coercion on the Rehabilitation of Prisoners *Douglas A. Bigelow and Richard H. Driscoll* 10
- The Development and Evaluation of Behaviorally Based Rating Scales *John P. Campbell, Marvin D. Dunnette, Richard D. Arvey, and Lowell V. Hellervik* 15
- Estimating the Influence of Job Information on Interviewer Agreement *John A. Langdale and Joseph Weitz* 23
- Leader Behavior Measurement in German Industry *D. Tscheulin* 28
- Effects of Different Leadership Styles on Group Accuracy *Joseph A. Cammalleri, Hal W. Hendrick, Wayne C. Pittman, Jr., Harry D. Blout, and Dirk C. Prather* 32
- Effects of Emotional Arousal on the Use of Supervised Coercion with Black and Union Employees *David Kipnis, Arnold Silverman, and Charles Copeland* 38
- The Influence of Sex-Role Stereotypes on Evaluation of Male and Female Supervisory Behavior *Benson Rosen and Thomas H. Jerdee* 44
- Relationship between Measures of Effort and Job Performance *William E. Williams and Dale A. Seiler* 49
- The Influence of Valence, Instrumentality, and Expectancy on Effort and Performance *Robert D. Pritchard and Mark S. Sanders* 55
- Expectancy Theory Predictions of Academic Effort and Performance *Terence R. Mitchell and Delbert M. Nebeker* 61
- Safety Training by Accident Simulation *Stanley Rubinsky and Nelson Smith* 68
- Information Seeking with Conflicting and Irrelevant Inputs *Jerrold Levine* 74
- An Experimental Investigation of Three Methods of Providing Weight and Price Information to Consumers *Robert D. Gatewood and Robert Perloff* 81

## SHORT NOTES

- A Test of the Need Gratification Theory of Job Satisfaction *James D. Neeley, Jr.* 86
- The Predictive Validity of Preliminary Performance Ratings by High School Personnel *James M. Orvik* 88
- An Application of Techniques To Shorten Tests and Increase Validity *John A. Fossum* 90
- Volunteer Satisfaction with In-Country Training for the Peace Corps: Reanalyses and Extended Findings *Richard R. Jones and W. J. Burns* 92

## LIST OF MANUSCRIPTS ACCEPTED

54



## INFORMATION FOR CONTRIBUTORS

**Style.** A professional article should possess certain characteristics: (a) conciseness and an apparent respect for reader time; (b) unambiguous and simple vocabulary with technical and erudite words used only when simpler ones would obviously be inadequate; (c) conformity to accepted technical style in tables, terminology, and references; (d) conclusions that are clearly related to the evidence presented. The reader should be led, step by step, from a statement of problem or purpose, through analysis of evidence, to conclusions and implications. Authors are encouraged to consult the little book entitled *The Elements of Style* by W. Strunk, Jr. and E. B. White (New York, Macmillan, 1962).

**Criteria for evaluation.** Manuscripts will be evaluated on the basis of several criteria, including: (a) significance in contributing new knowledge to the field, (b) technical adequacy, (c) appropriateness for the *Journal of Applied Psychology*, and (d) clarity of presentation.

**Format.** Manuscripts must be prepared in the format described in the *Publication Manual of the American Psychological Association* (1967 Revision), obtainable for \$1.50 from the American Psychological Association, 1200 17th Street, N.W., Washington, D.C. 20036. Articles not prepared in this manner cannot be reviewed. Special attention should be given to the section on typing the manuscript (p. 48) and to the sections on tables, figures, and references. Note that *all copy must be double spaced*, including references, title, figure captions, etc. The senior author's last name should appear on each page in the upper left-hand corner. In reference lists, give journal titles in full; do not abbreviate.

**Copies.** All manuscripts must be submitted in triplicate. One copy should be the original typed copy, the other two, clear carbon copies or photo reproductions, only. Authors should check the final typing carefully and retain a copy of the manuscript as a precaution against loss in the mail.

**Length.** Before preparing a manuscript, the author should check several recent issues to get an idea of the approximate length of regular articles published in the *Journal of Applied Psychology*. (One printed page equals roughly three double-spaced typewritten manuscript pages.) A few longer articles of special significance may be printed from time to time as monographs. Occasionally the *Journal* will have a section of "Short Notes" featuring brief reports on studies which make some methodological contribution or constitute an important replication.

**Abstracts.** Each copy of the manuscript must be accompanied by an abstract of 100-120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

**Figures.** The original drawing of a figure, or an 8 X 10-in. glossy print of the drawing, is needed for publication and must be submitted with the original typed copy of the manuscript. Duplicate copies of the figure may be photographic or pencil-drawn copies. Figures must be hand-lettered; typewritten lettering is not acceptable.

**Early publication.** Regular articles are published as nearly as possible in the order in which they were received. Although everything possible is done to keep publication lag to a minimum, at present there is a 10-12-mo. period between the time of receipt and publication of an article. The author of an accepted manuscript may secure early publication, usually in the first issue to go to press after its acceptance, by arranging for the article to be printed as extra pages, increasing the number of pages that subscribers receive. Information about charges for early publication will be sent with the letter of acceptance.

**Reprints.** No gratis reprints are supplied. Reprints may be ordered from the printer prior to publication.

**Supplementary material.** Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Authors should keep supporting and raw data for at least 5 yr. after publication of their article, and if applicable, offer them to the reader upon request.

**Copyrighted material.** The following policies govern reprinting or other forms of reproduction of materials copyrighted by the American Psychological Association: Written permission must be obtained for copying or reprinting any tables, figures, or text of more than 500 words in length. Persons or organizations who obtain permission to reproduce any such copyrighted materials must include on the first page of the reproduced materials the exact copyright notice which appeared on the Association's original publication. There will be a charge of \$10 per page for such permission except when the requester is the author of the article to be reprinted, or when the article is to be reproduced in a limited number of copies solely for instructional purposes.

## EFFECTS OF PERSONAL BACKGROUND AND TRAINING ON WORK VALUES OF THE HARD-CORE UNEMPLOYED<sup>1</sup>

JAMES G. GOODALE<sup>2</sup>

*Bowling Green State University*

This study described how work values of 110 disadvantaged persons differ from those of 180 unskilled and semiskilled employees, identified biographical correlates of work values, and examined changes in work values following training. When compared with regular employees, hard-core trainees placed less emphasis on the tendency to keep active on the job, taking pride in their work, and subscribing to the traditional Protestant Ethic, but placed more emphasis on making money on the job. Significant relationships were found between background characteristics and work values of the hard core. Changes in work values of disadvantaged subjects after 8 weeks of training did not differ from those of 252 controlled subjects (insurance agents and college students).

Persons classified as disadvantaged or hard core represent a subculture of our society with an indigenous life style and value system. One aspect of this value system that is of particular interest to social scientists is the concept of work values—an individual's attitude toward work in general rather than his feelings about a specific job. Many authors have speculated about the development of attitudes of the hard core, but they have presented few data to support their conclusions.

From a series of intensive interviews of 600 middle- and working-class families in Chicago, Davis (1946) identified three factors that may produce the behavior and set of values characteristic of the ghetto subculture. First, the necessity for survival forces the child of the lower-class family to seek immediate gratification of the most basic physical needs (food, clothing, and shelter), and it inhibits his striving for less urgent goals. Second, Davis argued that when a person becomes

accustomed to living at a subsistence level, unemployment becomes an acceptable norm. Third, his lack of adequate income, clothing, shelter, education, and vocational skills makes it impossible for the disadvantaged individual to escape the ghetto. In a similar essay, Himes (1968) observed that underprivileged black children who do not interact daily with employed persons fail to learn that effort leads to advancement in the work situation and remain naive about the language, dress, attitudes, and behavior expected by employers.

Unlike Himes, Schwartz and Henderson (1964) pointed out that most adolescents are exposed to the American work ethic through their experiences either at home or at school. They theorized that the disadvantaged are torn by the contrast between the ideals of the Protestant Ethic (Weber, 1958; e.g., work is good, achievement leads to advancement) and the reality of menial jobs, low pay, and chronic unemployment. They resolve this dilemma by devaluing work and by finding other ways of making money such as stealing, soliciting, and pushing dope. Their choice of solution reflects the rejection of legitimate employment as a means of advancement.

Despite the conclusions of the previous authors, Williams (1968) reasoned that the underprivileged accept the societal work ethic and want to support themselves through employment, but this desire is frustrated in demeaning, low-paying jobs. According to

<sup>1</sup> This research was supported under Grant 91-37-70-53 from the Manpower Administration, United States Department of Labor, under the authority of Title I of the Manpower Development and Training Act of 1962, as amended. The author wishes to thank Patricia C. Smith, O. W. Smith, J. P. Flanders, and A. G. Neal for helpful comments on earlier drafts of this article. Appreciation is also expressed to Allen Yates for his assistance in programming and analysis.

<sup>2</sup> Requests for reprints should be sent to James G. Goodale, now with the Faculty of Administrative Studies, York University, Downsview 463, Ontario.

Williams (1968) and Rainwater (1966), most hard-core males work for little money, and as this situation continues over a period of time, employment for low wages becomes aversive, although work itself is still valued. Williams (1968) claimed that since the disadvantaged do not differ from the rest of the labor force in their work values, a well-paying job will transform them into productive employees.

Some authors have measured values of the hard core. In an analysis of alienation scores, Bullough (1967) found that black residents of the ghetto expressed greater feelings of anomie and powerlessness than blacks living in integrated suburban areas. Agreeing with Davis (1946), Bullough concluded that work values of the hard core not only result from ghetto living but also perpetuate the impoverished environment.

Using a sample of disadvantaged persons, Wijting (1969) discovered relationships among work values and demographic information, parental models, early physical surroundings, and early psychological environment. In a canonical regression analysis, high incidence of police trouble in the family, rural residency, and low-family income were associated with emphasis on the social rewards of work and preference for being inactive and uninvolved on the job.

#### *Attempts To Hire the Disadvantaged*

Recognizing the vicious circle of unemployment experienced by members of the hard core, the federal government and private business launched a nationwide effort to hire and train the disadvantaged by creating a program named Job Opportunities in the Business Sector and an implementing agency known as the National Alliance of Businessmen (NAB). The NAB set as its goal the employment of 100,000 hard-core individuals by June 1969, and 500,000 by June 1971.

The NAB companies have made sincere efforts to hire hard-core applicants, to improve their skills in specialized training, and to place them on jobs requiring high levels of ability. However, the NAB program has not transformed all applicants into satisfied and productive workers. Of over 400,000 employees hired since 1968, 47% quit their jobs

within the first 6 months of employment.<sup>3</sup> In the metropolitan areas that have 100 or more companies participating in the NAB program, turnover rates vary greatly. For example, during the same period, one municipal area in New England reported a 20% turnover rate among the hard core, another in Wisconsin reported a 40% figure, and in Florida, a 56% figure was reported.<sup>4</sup>

High turnover, therefore, may involve the work values of employees. The NAB program has not dealt with these in an effective manner. The work values of disadvantaged employees seem to differ markedly from those held by all other workers in similar jobs, and, in addition, individual differences in attitudes toward work may exist among hard-core employees. In order to determine if these apparent differences are real, the values must be measured.

#### *The Current Study*

Although anecdotal evidence and turnover statistics have suggested that the disadvantaged appear to react to work situations differently than do other employees holding the same jobs, there is no research to explain how and why the two groups differ in their work values. This study, therefore, focused on work values of hard-core employees. Since the research was exploratory in nature, no formal hypotheses were formulated. The objectives of the project were as follows: (a) to measure the differences between work values of newly hired hard-core employees and those of other newly hired workers in similar jobs, (b) to identify background characteristics that are related to work values, and (c) to

<sup>3</sup> Figure presented by Paul W. Kayser, outgoing president of the NAB at the annual meeting in Washington, D.C., March 6, 1970.

<sup>4</sup> The fact that many people quit their jobs does not necessarily mean that the NAB program has failed or that the disadvantaged make unproductive or dissatisfied employees. Individuals may leave their jobs for reasons unrelated to the NAB program (e.g., to move to another city), but their turnover statistics would be included with those who left because they did not like the NAB program or because they did not like work. The statistics were included in the Confidential Progress Report issued by the NAB on January 31, 1970.

detect changes in work values as a function of orientation programs.

## METHOD

### Overview

The sample included subjects classified as disadvantaged<sup>5</sup> (hard-core group), regularly employed unskilled or semiskilled workers (comparison group), and middle-class persons (control group). To accomplish objective (a), work values were contrasted between the hard-core and comparison groups. Objective (b) concerned only the disadvantaged subjects. In meeting objective (c), changes in work values of the hard core were compared with those of the control group.

### Subjects

The group of disadvantaged subjects contained 37 females and 73 males, 99 who were black and 11 who were white. They ranged from 18 to 42 years of age with a mean of 23.5, and their educational level varied from 6 to 13 years with a mean of 10.6. The subjects averaged 4.5 years of previous work experience primarily in unskilled jobs. The comparison group included 139 semiskilled and unskilled employees of a midwestern glass-manufacturing company and 41 newly hired, hourly workers employed in a southern detergent factory. Serving as control subjects were 137 agents of an eastern insurance company and 115 undergraduates of a small college in California.

The 110 persons classified as hard core were selected from four companies affiliated with NAB. Only 13 subjects terminated before training was finished. Forty subjects hired by a plant in northeastern Ohio produced light bulbs. Thirty-five participants greased and assembled small parts in an automobile training center in northeastern Ohio. In a consortium of 16 companies in southern Ohio, 25 of the disadvantaged received training ranging from manual labor in a steel mill to clerical work in a local bank. Ten additional subjects performed general labor and material handling in a glass manufacturing factory in northwestern Ohio.

### Design

The design of this study can be categorized as a nonequivalent control group design (Campbell & Stanley, 1966) in which the control and experimental subjects are not randomly assigned to treatments. Nonequivalent subjects were used as a control group

because disadvantaged persons not involved in NAB training were unavailable. In addition, the study must be considered a quasiexperiment because training programs, which differed across companies, were regarded as the same treatment.

### Procedure

Participants were told their responses to questionnaires and interviews would provide information about differences in work attitudes, but would have no bearing on their jobs. Participation was voluntary, and subjects were assured that only general results would be reported to their employers. The investigator collected all data from hard-core persons, and questionnaires from other subjects were either mailed to their homes or administered by company personnel and then sent directly to the investigator.

Trainees spent approximately half of each work week in basic education and orientation and the other half in on-the-job training. Hard-core subjects completed the Survey of Work Values (SWV) shortly after they entered training (Time 1) and about 6 weeks later at the completion of the program (Time 2). This questionnaire can be scored on six subscales—Pride in Work, Job Involvement, Activity Preference, Attitude toward Earnings, Social Status of Job, and Upward Striving—and on six clusters—Intrinsic Work Values, Organization-Man Ethic, Upward Striving, Social Status on Job, Conventional Ethic, and Attitude toward Earnings (see Wollack, Goodale, Wijting, & Smith, 1971, for definitions). In the development of the SWV, industrial employees assigned items to their respective subscales with high reliability. When using the scale, subjects are instructed to agree or disagree with each of 54 statements. Scores are obtained by summing responses to items comprising each of six subscales. The test-retest reliabilities of the 9-item subscales range from .68 to .76 despite the fact that the items have been chosen to vary in endorsement level (Wollack et al., 1971). The reading level of the SWV is low enough to permit its use with disadvantaged applicants (Wijting, 1969).

By filling out a biographical inventory at Time 1, each subject supplied information about the physical and psychological conditions of early home life, the presence of parental work models in the home, the area of the country and size of city in which the person was raised, his work experience, educational and occupational level, financial responsibility, and recent work record. At Time 2, hard-core employees discussed the experiences that were especially satisfying or dissatisfying to them during training in an interview with the investigator.

Comparison employees completed the SWV only once, either shortly after being hired or after an unrecorded amount of experience on the job. Control subjects responded to the SWV once and then a second time about 2 months later. They continued in their usual school or work activities between administrations of the SWV.

<sup>5</sup> A person who is classified as disadvantaged must be a member of a poor family and be unemployed, underemployed, or hindered from seeking work and be at least one of the following: (a) school dropout, (b) minority member, (c) under 22 years of age, (d) 45 years of age or over, and (e) handicapped (Ohio Bureau of Employment Services Letter No. 1055, March 19, 1969).

TABLE 1  
DISCRIMINANT FUNCTION ANALYSIS WITH SWV SUBSCALES AS CRITERIA

Subscale	Hard core (N = 110)		Comparison (N = 180)		Discriminant function		
	M	SD	M	SD	b <sup>a</sup>	s <sup>b</sup>	Contri- butio- n
Social Status of Job	12.53	2.24	12.89	1.87	.209	.170	.036
Activity Preference	15.85	1.88	17.07	.94	.705	.772	.409
Job Involvement	16.53	1.63	17.03	1.11	-.048	.347	.011
Upward Striving	15.66	1.83	15.87	1.45	.129	.119	.012
Attitude toward Earnings	14.03	2.27	12.39	1.99	-.625	-.685	.484
Pride in Work	16.99	1.63	17.64	.65	.223	.523	.068

Note. Wilks' lambda = .734,  $\chi^2 = 104.06$ ,  $p < .001$ .

<sup>a</sup> Regression weight.

<sup>b</sup> Correlation of variate with its composite.

### Analysis

All analyses involving SWV responses were performed separately for subscale scores and cluster scores. Absences and incomplete questionnaires created the problem of missing data, often encountered in field research. When cluster or subscale scores were computed, the mean of available responses to items of that cluster or subscale was inserted for missing values (Timm, 1970).

### RESULTS

#### Comparison of Work Values

Discriminant function analysis revealed differences between work values of the hard-core and comparison groups. This statistical technique determined the weighted combination of SWV scores discriminating maximally between the two groups of subjects (Cooley & Lohnes, 1962). The correlations of each variable with the discriminant function (Kelly, Beggs, McNeil, Eichelberger, & Lyon, 1969) and their contribution to the unit variance of the discriminant function indicated the dimensions of work values on which the two groups differed most.

The composite of subscale scores (see Table 1) discriminated between the hard-core and comparison groups with a  $\chi^2$  of 104.06 ( $df = 6$ ,  $p < .001$ ). Attitude toward Earnings contributed .48 to the unit variance of the discriminant function, while Activity Preference and Pride in Work accounted for .41 and .07, respectively. Therefore, the main

contrast between the two groups was in their preference for activity and deemphasis of money; the hard-core persons scored 5.88 on the discriminant axis, and the regular employees scored 7.12. The subjects were very similar, however, in Job Involvement, Upward Striving, and Social Status of Job.

Analysis of SWV cluster scores (see Table 2) also produced highly significant differentiation between the two groups of employees ( $\chi^2 = 81.47$ ,  $df = 6$ ,  $p < .001$ ). Since Conventional Ethic and Upward Striving correlated negatively with the discriminant function, but Attitude toward Earnings correlated positively, the composite reflected an emphasis on wages and a deemphasis of the conventional work ethic. Hard-core subjects scored 7.39 on the composite, while the comparison employees scored 6.30. The two groups were comparable in Social Status of Job, Organization-man Ethic, and Intrinsic Work Values.

#### Correlates of Work Values

Next, the relationships among personal background variables and work values of the hard core were investigated. Since it was decided to combine the biographical information into more reliable and interpretable variables, the 28 background items were subjected to a principal components factor analysis with varimax rotation. After seven factors

TABLE 2  
DISCRIMINANT FUNCTION ANALYSIS WITH SWV CLUSTERS AS CRITERIA

Cluster	Hard core (N = 110)		Comparison (N = 180)		Discriminant function		
	M	SD	M	SD	b <sup>a</sup>	s <sup>b</sup>	Contri- butio- n
Intrinsic Work Values	23.06	2.21	23.69	1.17	-.052	-.395	.022
Organization-man Ethic	16.70	2.09	16.97	1.30	-.081	-.171	.015
Upward Striving	8.64	1.40	8.97	.90	-.335	-.295	.075
Social Status of Job	10.25	1.86	10.39	1.64	-.217	-.089	.022
Conventional Ethic	20.39	1.84	21.08	.99	-.403	-.505	.187
Attitude toward Earnings	9.36	1.70	8.13	1.55	.818	.744	.678

Note. Wilks' lambda = .779,  $\chi^2 = 81.47$ ,  $p < .001$ .

<sup>a</sup> Regression weight.

<sup>b</sup> Correlation of variate with its composite.

were extracted, less than 5% of the residual correlations were greater than .10.

Factor 1 (Economic Maturity) was defined by positive loadings on age, years of work experience, marital status, number of persons supported, and by a negative loading on person paying the bills. An individual scoring high on this factor was likely to be old, married with several dependents, to pay most of his family's bills, and to have had much work experience. Factor 2 (Police Trouble) had positive loadings on items dealing with frequency and severity of police trouble by members of one's family and number of arguments with one's parents. Factor 3 (Rural South-Urban North) correlated with the area of the country and size of city in which a person spent his early life. A high score on Factor 4 (Welfare) represented an individual whose father was often out of work and whose family was on welfare. Factor 5 (Socioeconomic Status) summarized the educational level of one's parents and number of family members who drank excessively. Factors 6 and 7 were poorly defined and were not included in subsequent analyses.

Values of the 15 items composing the interpretable factors were converted to  $z$  scores and summed to form five clusters. These variates, along with sex and educational level, were included in canonical regression analyses as predictors of SWV subscale and cluster scores. Eleven retrospective variables were

dropped because of low variance or high percentage of missing data.

Canonical regression analysis determined the linear combination of the set of predictors and the set of criteria that maximized the correlation between the two sets of variates (Bartlett, 1941; Burt, 1948; Horst, 1961). The correlation of each variate with its composite (Meredith, 1964), and the contribution of a given variate to the unit variance of its composite were used to interpret the canonical analysis. The following interpretations must be considered tentative since cross-validation of the canonical correlations was not feasible because of the small number of subjects. Similarity of the present results to those of previous studies, however, added credibility to the relationships described below.

The first canonical correlation between background variates and SWV subscale scores was .612 ( $\chi^2 = 64.33$ ,  $df = 42$ ,  $p < .025$ ). The predictor composite (see Table 3) showed positive loadings on Economic Maturity (.477), Educational Level (.585), and Rural South-Urban North (.231) and a negative loading on Welfare (-.620). The criterion composite correlated positively with Job Involvement (.880) and Pride in Work (.469) and negatively with Social Status of Job (-.419). The predictor composite described a person from the urban North, who was relatively well educated and economically

TABLE 3  
CANONICAL ANALYSIS WITH SWV SUBSCALES  
AS CRITERIA ( $n = 78$ )

Variate	b <sup>a</sup>	s <sup>b</sup>	Contribution
Economic Maturity	397	477	189
Police Trouble	002	046	000
Rural South-Urban North	500	231	116
Welfare	-483	-620	299
Socioeconomic Status	232	239	056
Sex	307	103	032
Educational Level	527	585	308
Social Status of Job	-349	-419	146
Activity Preference	094	452	042
Job Involvement	786	880	691
Upward Striving	-089	268	-024
Attitude toward Earnings	-121	-423	051
Pride in Work	197	469	093

Note. Sample included only hard-core subjects. Decimal points are omitted.

<sup>a</sup> Regression weight.

<sup>b</sup> Correlation of variate with its composite.

mature, and whose family had spent little or no time on welfare. This type of person values being highly involved in his job and taking pride in his work but deemphasizes the social status of being employed.

The analysis of SWV cluster scores and biographical data ( $Rc = .572$ ,  $\chi^2 = 61.53$ ,  $df = 42$ ,  $p < .025$ ) produced very similar results. The predictor composite in Table 4, composed of Educational Level (.681), Economic Maturity (.406), and Welfare (-.432), described a person of relatively high educational level and economic maturity whose family had spent little or no time on welfare. The criterion function showed positive loadings on Intrinsic Work Values (.724) and Conventional Ethic (.623) and a negative loading on Social Status of Job (-.558). This composite represented a person who values work as its own reward and deemphasizes the social status of being employed.

#### Modification of Work Values

The next analysis tested the significance of changes in work values experienced by hard-core subjects. Only 65 disadvantaged persons filled out the SWV at Time 2 because of absences, terminations, and refusal of several

subjects to take a questionnaire twice in 2 months. Differences were computed by subtracting SWV scores of Time 1 from those of Time 2. Changes in work values of the hard-core subjects ranged from -.25 to .40 and did not differ significantly from those of the control group.

#### Subjects' Impressions of NAB

Shortly before completion of the orientation, subjects were asked their feelings about the program and what training experiences they found especially satisfying and dissatisfying. Originally, a content analysis of these responses was planned, and frequency of response was to be correlated with changes in work values. This step was dropped when no significant changes in work values were found, but subjects' impressions were still informative. Over 90% said the program was helpful because it provided them with an opportunity to work and earn money. Many with poor work records viewed the training as a second chance to secure gainful employment. Most frequently mentioned as dissatisfying were routine, low-level work, poor condition of training materials, and close supervision by company personnel.

TABLE 4  
CANONICAL ANALYSIS WITH SWV CLUSTERS  
AS CRITERIA ( $N = 78$ )

Variate	b <sup>a</sup>	s <sup>b</sup>	Contribution
Economic Maturity	459	406	186
Police Trouble	-032	-056	002
Rural South-Urban North	465	222	103
Welfare	-248	-432	107
Socioeconomic Status	128	150	019
Sex	441	263	466
Educational Level	685	681	325
Intrinsic Work Values	449	724	009
Organization-man Ethic	092	100	071
Upward Striving	166	430	290
Social Status of Job	-520	-558	283
Conventional Ethic	454	623	022
Attitude toward Earnings	-084	-252	

Note. Sample included only hard-core subjects. Decimal points are omitted.

<sup>a</sup> Regression weight.

<sup>b</sup> Correlation of variate with its composite.

## DISCUSSION

### *Work Values of the Disadvantaged*

Results of the two discriminant function analyses strongly supported the premise that the hard core differ markedly from regular employees in their expressed work values. The hard-core subjects scored lower than the comparison group in Activity Preference, Pride in Work, Upward Striving, and Conventional Work Ethic and higher in Attitude toward Earnings. These data indicated that the disadvantaged labor primarily for money rather than for the intrinsic rewards of work. Davis (1946), Himes (1968), and Schwartz and Henderson (1964) also noted the tendency of the hard core to concentrate on immediate gratification and to devalue work for its own sake. Bullough (1967), Killian and Grigg (1962), and Lefton (1968) made similar conclusions because ghetto blacks expressed greater feelings of alienation from the traditional work ethic than did whites or well-to-do blacks. Also supporting this general trend were Centers' report (1949) that lower-class groups strongly valued security and money and Bloom and Barry's finding (1967) that blacks emphasized extrinsic work rewards more than did whites.

The canonical analyses disclosed some important variation in work values within the hard-core sample. Disadvantaged persons of relatively high educational level and economic maturity showed positive attitudes toward the Conventional Work Ethic and the intrinsic rewards of work (Pride in Work and Job Involvement) and placed less emphasis on the social status of employment. Goodale (1970) found that individuals of high socioeconomic status also subscribed to the conventional work ethic. It is interesting to note that Attitude toward Earnings, the work value that discriminated most significantly between hard-core and employed persons, was not related to biographical characteristics of the hard-core sample.

The canonical analyses revealed correlates of work values alien to those held by working members of society. However, longitudinal studies that trace work values developing in children of various socioeconomic classes are needed to identify the time at which the value

systems diverge and to suggest determinants of different sets of work values. Until developmental research is done, studies of work values and background information will be more descriptive than explanatory.

### *Changes in Work Values*

An examination of NAB programs that stressed attitude change would have been preferred, but such programs were not available. Perhaps because the emphasis was on acquisition of skill and educational improvement rather than on attitudes, the work values of hard-core subjects were not significantly altered by orientation. Outlines of the training schedules documented that little time was spent on attempts to modify work values of the participants.

It is unlikely that 8 weeks of training could have changed work values that have been formed by many years of experience. A reason for this may be that disadvantaged persons received training for routine, unstimulating jobs, while being told that they should regard work as intrinsically rewarding. The hard core may become disillusioned with their jobs when expectations formed in training are not fulfilled. Supporting this speculation is the finding of Quinn, Fine, and Levitin (1970) that the disadvantaged gave poor working conditions most frequently as the reason for quitting NAB jobs.

Several speculations can be made regarding methods of training that are likely to produce changes in work values. First, since this study disclosed specific work values in which hard-core and regular employees differed, NAB orientation could focus on altering those values. Second, the variance in work values within the disadvantaged sample indicated the necessity of having training tailored to individual needs. Counselors with information about a person's background and initial work values could develop personalized plans for training. Third, subjects could be allowed to move in sequential progression toward completion of their training instead of having to remain in the program for a fixed amount of time. Fourth, trainees could be placed on jobs after consideration has been given to abilities and successes demonstrated in training as well

as to the availability of jobs. These suggestions are made as alternatives to be tried and evaluated, not as rigid guidelines for successful NAB programs.

### *A Problem of Measurement*

Measurement of values is difficult when subjects are aware of socially desirable responses. It is legitimate to ask if the differences in work values of hard-core and comparison subjects were partially due to the desire of regular employees to gain approval with their SWV responses. If this is the case, why were the disadvantaged unconcerned with how their responses would appear? Davis (1946), Himes (1968), and Schwartz and Henderson (1964) posit that the disadvantaged are not cognizant of socially acceptable work values because of their isolated work subculture, and, therefore, cannot pretend to subscribe to them. Williams (1968) would argue, however, that disadvantaged subjects are aware of but do not endorse the prevailing work ethic because their current work situations contradict it.<sup>6</sup> Williams (1968) hypothesized that a hard-core trainee would accept the Protestant Ethic only if he were given a good job.

### *Conclusions and Implications for Future Research*

Although no changes in work values were detected immediately after orientation in this study, the NAB program may still alter attitudes. Work values of trainees should be measured several months after they have begun their jobs to see if they have accepted a new orientation toward employment. It appears that the hard core are more likely to improve both work values and performance after they have had some experience with jobs more closely matched to their abilities and interests.

Despite problems of measurement, this study gave more precise information regarding the work values of the hard core and ordinary employees in comparable jobs and

identified background characteristics that might have produced differences between the two groups. Unfortunately, comparison of performance of hard-core and regular employees on the job was impossible because current high unemployment prevented trainees from moving into full-time work. Relationships between work values and job performance should be examined, however, to discover how different orientations toward work correlate with performance on the job.

### REFERENCES

- BARTLETT, M. S. The statistical significance of canonical correlation. *Biometrika*, 1941, 32, 29-37.
- BLOOM, R., & BARRY, J. R. Determinants of work attitudes among Negroes. *Journal of Applied Psychology*, 1967, 51, 291-294.
- BULLOUGH, B. Alienation in the ghetto. *American Journal of Sociology*, 1967, 72, 469-478.
- BURT, C. Factor analysis and canonical correlations. *British Journal of Statistical and Mathematical Psychology*, 1948, 1, 95-106.
- CAMPBELL, D. T., & STANLEY, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally, 1966.
- CENTERS, R. *The psychology of social class*. Princeton, N.J.: Princeton University Press, 1949.
- COOLEY, W. W., & LOHNESS, P. R. *Multivariate procedures for the behavioral sciences*. New York: Wiley, 1962.
- DAVIS, A. The motivation of the underprivileged worker. In W. F. Whyte (Ed.), *Industry and society*. New York: McGraw-Hill, 1946.
- GOODALE, J. G. A canonical regression analysis of work values and biographical information of civil service employees. Unpublished manuscript, Bowling Green State University, 1970.
- HIMES, J. Work values of Negroes. In L. A. Ferman, J. L. Kornbluh, & J. A. Miller (Eds.), *Negroes and jobs*. Ann Arbor: University of Michigan Press, 1968.
- HORST, P. Relations among  $m$  sets of measures. *Psychometrika*, 1961, 26, 129-149.
- KELLY, F. J., BEGGS, D. L., MCNEIL, K. A., EICHELBURG, T., & LYON, J. *Research design in the behavioral sciences: Multiple regression approach*. Carbondale and London: Southern Illinois University Press and Fesler & Simons, 1969.
- KILLIAN, L. M., & GRIGG, C. M. Urbanism, race, and anomia. *American Journal of Sociology*, 1962, 67, 661-665.
- LEFTON, M. Race, expectations, and anomia. *Social Forces* 1968, 46, 347-352.
- MEREDITH, W. Canonical correlations with fallible data. *Psychometrika*, 1964, 29, 55-65.
- QUINN, R. P., FINE, B. D., & LEVITIN, T. Turnover and training: A social-psychological study of disadvantaged workers. Ann Arbor, Mich.: Author 1970. (Mimeo)

<sup>6</sup> A simple way to test whether the hard core are aware of socially acceptable work values would be to instruct them to fill out the SWV as they think a white-collar employee would.

- RAINWATER, L. Crucible of identity: The Negro lower-class family. *Daedalus*, 1966, 95, 172-211.
- SCHWARTZ, M., & HENDERSON, G. The culture of unemployment: Some notes on Negro children. In A. B. Shostak & W. Comberg (Eds.), *Blue-collar world: Studies of the American worker*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.
- TIMM, N. H. The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika*, 1970, 35, 417-437.
- WEBER, M. *The Protestant Ethic and the spirit of capitalism*. New York: Scribner, 1958.
- WIJTING, J. P. A canonical regression analysis of background variables and work values of underprivileged workers in Toledo, Ohio. Unpublished manuscript, Bowling Green State University, 1969.
- WILLIAMS, W. Manpower problems in the hard-core ghetto. In N. A. Palombra & E. B. Jakubauskas (Eds.), *An Interdisciplinary approach to manpower research*. Ames, Iowa: Industrial Relations Center, Iowa State University, 1968.
- WOLLACK, S., GOODALE, J. G., WIJTING, J. P., & Smith, P. C. The development of the Survey of Work Values. *Journal of Applied Psychology*, 1971, 55, 331-338.

(Received August 9, 1971)

## EFFECT OF MINIMIZING COERCION ON THE REHABILITATION OF PRISONERS<sup>1</sup>

DOUGLAS A. BIGELOW<sup>2</sup> AND RICHARD H. DRISCOLL

*University of Colorado*

Etzioni's theory of power and involvement in organizations formed the basis for an approach to the process of rehabilitation. A cross-sectional design was employed to test five hypotheses derived from this theoretical formulation. The analysis compared two inmate groups, in a federal youth correctional center, which differed in the degree to which they were subject to coercive power by the staff. It was found that coercive power was inversely associated with (a) cooperative attitudes among inmates, (b) normative expectations and pressures for cooperation with the supervisors, (c) a cooperatively disposed informal inmate leadership, and (d) a perception of the supervisors as having socially adaptive work values. These were believed important in rehabilitation. The final hypothesis—that noncoerced subjects would have adopted socially adaptive work values—was found to be in the expected direction but not significant.

The socialization of persons in an institutional setting is influenced by a number of aspects of the institution as an organizational structure. The present study is concerned with one of these aspects: the nature of power exercised by institutional authorities, as a factor bearing on the rehabilitation of young men in a federal correctional center.

Our conceptual analysis of institutions is based on that of Etzioni (1961, 1968) in which three categories of variables were described: status, power, and involvement. The members of an organization were divided into two statuses: the *elite* and the *rank and file*. There were three modes of power by which the elite might relate to the rank and file: *coercive*, *normative*, and *remunerative*. Coercive power consists of the use or threatened use of force to limit the behavioral alternatives of an individual, whereas normative power consists of persuasion, suggestion, and the use of interpersonal rewards to influence the choice of alternatives. Etzioni conceptualized coercive and normative power as tending to be mutually exclusive. Power, therefore, may be considered as lying along a single dimension—from high to low coercive power.

<sup>1</sup> The authors are indebted to William A. Scott for his supervision of this study and his constructive criticism of the manuscript.

<sup>2</sup> Requests for reprints should be sent to Douglas A. Bigelow, Department of Psychology, University of Colorado, Boulder, Colorado 80302.

(Remunerative power is not dealt with in the present study.)

The third variable in Etzioni's conceptualization was the involvement of the rank and file in the organization, which may range from *alienation* to *commitment*. To be alienated from the organization and its elite is to have an uncooperative attitude, to be disposed to resist directives, to antagonize, and to reject the authority, credibility, and influence of the elite. To be committed is to have a cooperative attitude, to be willing to obey, to please, and to learn from the elite. In this study, involvement is conceptualized as a single dimension—running from low commitment to high—and is specified as (a) cooperative orientation, (b) perception of the elite as possessing socially adaptive values, and (c) commitment to the general objectives of the institution.

Etzioni argued that, where the elite exercises a predominantly normative power over the rank and file, the latter tends to place itself into a cooperative relationship with the elite, committing itself more completely to the directives of the elite and to the goals and practices of the organization. On the other hand, where the elite exercises coercive power, the rank and file tends to be alienated. The central concern of this study is a specification, in the setting of a correctional institution, of Etzioni's general proposition about this relationship of power and involve-

ment. The investigators propose that an inmate population which interacts with a supervising staff exercising predominantly noncoercive (normative) power tends to respond cooperatively and to perceive the supervisory staff as worthy models and, further, that effective resocialization of the inmates occurs under these conditions.

The commitment of the rank and file is lodged in individual attitudes, in the normative expectations and pressures of the group, and in the informal leadership of the group.

For organizations whose function is the inculcation of socially adaptive values, the extent to which the rank and file adopts those values is the ostensible criterion of the institution's success; and, in our theoretical formulation, it is the outcome of day-to-day cooperative attitudes, group pressures, and modeling of the elite—which are associated with the exercise of minimal coercion on the part of the supervising staff.

The theoretical formulation presented above may be stated as five hypotheses. These hypotheses were made with respect to segments of an inmate population, each under different supervisory staff, in a federal youth correctional center (a more complete description appears below). In addition, the design of the study requires that the five groups be significantly different in the degree of coercive power exercised by their elite. The hypotheses were as follows: (a) Members of the coerced groups have less cooperative attitudes than those of the noncoerced groups toward their respective supervisors. (b) The coerced groups have less cooperative normative expectations. (c) They choose less cooperatively oriented leaderships than do the noncoerced groups. Finally, (d) the noncoerced groups perceive their supervisors as holding what is described below as better work values and (e) they have, themselves, assimilated better work values than have coerced groups of inmates.

#### METHOD

##### *Subjects and Institution*

The program of this rehabilitational institution involved academic education, vocational training in workshops, recreation, and a variety of other activities such as kitchen duties and grounds maintenance.

On the basis of discussion with staff and inmates, it was decided that the workshops were the locus of predominantly noncoercive relationships, while certain dormitories were the locus of predominantly coercive relationships.

Of a total of 97 subjects, 59 were in the coerced groups from the dormitory and 38 were in the non-coerced group from four workshops. All were in their late adolescence and were inmates of the federal correctional institution, having committed more than minor offenses.

#### *Measures*

A questionnaire consisting of six scales was administered to the dormitory and the shop groups. The dormitory form referred to "dormitory officers" and the shop form to "shop foremen."

*Coercive Power scale.* This scale assessed the extent to which the subjects' supervisors were seen as using force or threatened force to control inmates' behavior. There were six items, half of them reverse scored, for example, "A guy has to be on his toes to keep from getting into trouble with the shop foreman" and "The shop foreman doesn't worry about inmates breaking minor rules, and doesn't come down too hard on them for it." Response alternatives were "agree" or "disagree," with those indicating coercion scored high.

*Cooperative Attitude scale.* It was hypothesized that inmates not subject to the exercise of coercive power would tend to have more cooperative attitudes. Of the six agree-disagree items half were reverse scored, for example, "If the shop foreman asked me to clean up the shop right away, I would do it, even if he wasn't around to see that I did" and "I try to get away without doing things the shop foreman asks me to do whenever I can."

*Cooperative Norm scale.* This scale assessed the subjects' perception of their group's normative expectations and pressures with respect to cooperation with the supervising elite. There were six agree-disagree items, half reverse scored, for example, "The guys in this group think that the shop foreman only asks us to do what is fair and reasonable" and "If you want to get along well with the guys in this group, you can't be too friendly with the shop foreman."

*Work Values of the Elite scale.* Learning to be an acceptable member of society, for members of this socioeconomic group, consists largely in learning to be a motivated, stable worker who believes that worthwhile rewards, that is, prestige, personal satisfaction, and security, as well as remuneration result from vocational diligence and performance. Two of the eight items were: "I think the shop foreman's family and friends respect him because he holds a steady job" and "I am sure the shop foreman finds his job dull and boring; I can't imagine why he has stayed at it this long." The scale had three response alternatives: "agree," "disagree," and "don't know." Good work values were scored high.

*Work Values of Subject scale.* The ultimate objective of the institution and, therefore, of the exercise

TABLE 1  
PSYCHOMETRIC PROPERTIES OF THE SCALES

Scale	1	2	3	4	5	6
Coercive Power (1)	(.69)					
Cooperative Attitude (2)	-.49*	(.68)				
Cooperative Norms (3)	-.37*	.46*	(.59)			
Work Values of Elite (4)	-.64*	.49*	.35*	(.80)		
Work Values of Subject (5)	-.19*	.33*	.10	.25*	(.61)	
Leadership (6)	-	-	-			(.84)
Homogeneity ratios	.27	.27	.20	.34	.19	.57

Note.  $N = 97$ ;  $r \geq .17$  is required for  $p < .05$ , one-tailed test. Cronbach's alphas are in parentheses.

\*  $p < .05$ .

of power is the rehabilitation of the inmates: Either they learn socially adaptive values and habits or the institution has failed. In order for the released inmate to successfully adjust to society, it is essential that any positive values he may have been exposed to in the correctional center be assimilated and become an enduring, personal commitment to a socially viable life-style. This scale measured the extent to which a steady job, self-improvement, and productivity were important as personal standards. Unlike the above scales, this one measured values and attitudes that were general orientations and not related to specific circumstances within the institution. Several items forced choices between this commitment and more exciting alternatives—the kind of choices that the released inmate would have to make. Two of the seven items were: "Getting an education or training is worth it to me in the long run—even if it means having less fun and fewer friends right now" and "Almost all jobs are dull and boring; it's no wonder a guy can't stay on the job for very long." There were two response alternatives; good work values were scored high.

*Leadership scale.* This scale was designed with the assumption that leadership may be considered to be distributed among the members of the group. Over time and across situations every group member is more or less a leader. Each subject made up to four ranked choices from among his fellow group members per item; each choice being weighted according to its rank. The items were: "In this group of people, which one(s) would you most likely listen to, if he made a suggestion to you and your friends?" "Which person would you want to speak for you, if your group had to talk with the shop foreman about something important?" "Which person in this group really knows what is going on around here?" "Which person is the most popular in the group?" A subject's scores consisted of the weighted frequency of his being chosen on these items.

## RESULTS

### *Validity of Scales and Constructs*

Scott's (1960) homogeneity ratios, reliabilities (Cronbach's alphas), and intercor-

relations of scales are shown in Table 1. The homogeneity ratios were all within the acceptable range of from .15 to .60, between the boundaries of minimum coherence and maximum redundancy. The scale reliabilities were all higher than the scale intercorrelations, indicating discriminant validity of the scales.

It was expected that low coercive power by the elite would be associated with cooperative attitudes, normative expectations for cooperation, and a perception of the elite as worthy models. As seen in Table 1, the scales measuring the latter three constructs correlated negatively with the Coercive Power scale, which bears out the theoretical expectation and indicates the validity of the constructs. It was also expected that work values of the subject are associated with these factors; this was also supported by the pattern of correlations.

Table 2 presents the correlations among scales for the coerced and noncoerced subject groups, separately. These correlations are lower than for the two groups combined, but are still strong in the expected directions. Minor depression of the correlations can be attributed to reduced response variance within the groups, while the remaining similarity in the pattern of correlations indicates that they are not due primarily to situational differences.

### *Tests of the Hypotheses*

The methodological presupposition was that the elite supervising the dorm groups of inmates would be seen as exercising more coercion than would the elite supervising the

TABLE 2  
INTERCORRELATIONS BETWEEN SCALES FOR COERCED AND NONCOERCED SUBJECTS SEPARATELY

Scale	CP*	CA	CN	WVE
Coerced group				
Cooperative Attitude	-.39*			
Cooperative Norms	-.34*	.50*		
Work Values of Elite	-.55*	.41*	.34*	
Work Values of Subject	-.15	.31*	.04	.20
Noncoerced group				
Cooperative Attitude	-.22			
Cooperative Norms	-.34*	.33*		
Work Values of Elite	-.53*	.11	.30*	
Work Values of Subject	-.18	.14	.14	.29*

\* Coercive Power scale.

\*  $p \leq .05$ .

workshop groups. This was a test of the design of the study to ensure that the designation of subjects as coerced or not coerced was, in fact, correct. The comparison of mean scores on the Coercive Power scale, presented in Table 3, support this designation.

The first hypothesis was that inmate groups interacting with a less coercive elite would have more cooperative attitudes than would inmate groups interacting with a more coercive elite. Mean scores on the Cooperative Attitude scale presented in Table 3 support this hypothesis, as well as the similar second hypothesis which predicted a relationship between coercion and normative expectations for cooperation with the elite.

The third hypothesis was that coerced inmate groups would evolve a less cooperatively oriented leadership. It was expected, then, that among coerced subjects the Leadership rating of a subject would be negatively

correlated with his self-reported Cooperative Attitude score, but positively correlated among noncoerced subjects. For the coerced subjects the correlation was  $-.20$  and for the noncoerced,  $.24$ . The difference between the correlations was significant in the direction predicted at the  $p < .025$  level, which supports the third hypothesis. A more coercive elite is associated with a less cooperative rank and file leadership, as well as with less cooperative group norms and individuals with less cooperative attitudes.

The fourth hypothesis was that the more coerced subjects would perceive their supervisors as having less desirable work values. The mean scores on the Work Values of the Elite scale is presented in Table 3. These data support the fourth hypothesis: The less coercive elite is perceived as a better model for the learning of socially adaptive work values.

TABLE 3  
COMPARISON OF MEAN SCORES BETWEEN COERCED AND NONCOERCED GROUPS

Scale	Scale range	Coerced subjects	Noncoerced subjects	t	p≤
Coercive Power	6	2.84	1.26	5.16	.001
Cooperative Attitude	6	4.26	5.53	5.30	.001
Cooperative Norms	6	3.79	4.47	2.10	.025
Work Values of Elite	16	7.91	11.37	5.14	.001
Work Values of Subject	7	5.35	5.74	1.17	.150

The fifth hypothesis was that the less coerced subjects would have assimilated (or be in the process of assimilating) more socially adaptive work values than would coerced subjects. The mean scores on the Work Values of Subjects scale, presented in Table 3, lend only suggestive support to this hypothesis. The difference is in the expected direction, but is not significant.

### DISCUSSION

The data support most of our theoretical formulations. The exercise of coercive power by the elite was associated with alienation of the rank and file, which pervades individual attitudes, group normative expectations, and group leadership. This alienation included not only the noncooperative disposition of the rank and file, but also their perception of the elite as negative role models. Conversely, low coercive power was associated with cooperative dispositions and positive perception of the elite, factors commonly believed to be important in the learning of positive work values, and thus in the rehabilitation of inmates. Cooperative individual attitudes inside the prison can lead to a more general orientation toward cooperation with authorities and employers upon release. Cooperatively oriented group pressures and interpersonal pay-offs can influence individual attitudes and behaviors in the direction of the group norms. Finally, perception of significant others as positive role models can aid in the learning of adaptive values and habits.

The objective of the correctional institution studied here was the rehabilitation of inmates. While our data showed that several rehabilitative factors were associated with a lack of coercion, they did not indicate that inmates from the noncoerced group were further along the rehabilitative process as measured by their positive work values. The failure to find better work values among the noncoerced group can be explained by their

insufficient experience in noncoercive settings. Inmates in noncoercive groups were often in more coercive groups at other times in their daily routine, and the amount of time spent in any one group was limited (by rotations, releases from prison, etc.). Rehabilitation must be seen as a protracted process, with the influences being applied consistently over a significant period of time before appreciable changes in values are made.

Since the design of this study is not longitudinal, any trend of value change in the inmate is not revealed. Further, no cause-effect relationship between the exercise of coercive power and involvement can be demonstrated. Despite these limitations, the close relationship between the lack of coercive power, inmate involvement, and rehabilitation does support the use of noncoercive power by the elite. It seems clear that an institution may more effectively pursue its goals by building on normative relationships between the rank and file than by the use of coercive power. In the correctional institution studied, it was possible to build normative relations around the mutually challenging and intrinsically motivating tasks of the workshop situation.

Similar studies in other institutions would be needed to provide a sound basis for generalizing the present findings: Commitment of the rank and file, which seems necessary to the accomplishment of organizational goals, is associated with the exercise of normative rather than coercive power by the organizational elite.

### REFERENCES

- ETZIONI, A. *A comparative analysis of complex organizations*. New York: Free Press, 1961.
- ETZIONI, A. *The active society: A theory of societal and political processes*. New York: Free Press, 1968.
- SCOTT, W. Measures of test homogeneity. *Educational Psychological Measurement*, 1960, 20, 751-757.

(Received July 19, 1971)

# THE DEVELOPMENT AND EVALUATION OF BEHAVIORALLY BASED RATING SCALES

JOHN P. CAMPBELL<sup>1</sup>

*University of Minnesota*

MARVIN D. DUNNETTE AND RICHARD D. ARVEY<sup>2</sup>

*University of Minnesota and Personnel Decisions, Inc., Minneapolis*

LOWELL V. HELLERVIK

*Personnel Decisions, Inc., Minneapolis*

A distinction is made between criterion measures that assess individual performance in terms of concrete job functions and those that reflect organizational outcomes several steps removed from actual behavior (e.g., salary level). It is argued that psychologists should be trying to measure and predict the former, and a modification of the method of scaled expectations is suggested as one technique for doing so. The method was used to develop nine criterion dimensions for department managers in a nationwide retail chain. The resulting behavior rating scales were compared with a summated ratings technique on a sample of 537 department managers. The behavioral scales yielded less method variance, less halo error, and less leniency error. Additional benefits from the method are also noted.

Campbell, Dunnette, Lawler, and Weick (1970) have distinguished among the concepts of behavior, performance, and effectiveness as three outcomes of organizational roles. *Behavior* is simply what people do in the course of working (e.g., dictating letters, giving directions, sweeping the floor, etc.). *Performance* is behavior that has been evaluated (i.e., measured) in terms of its contribution to the goals of the organization. Finally, *effectiveness* refers to some summary index of organizational outcomes for which an individual is at least partially responsible such as unit profit, unit turnover, amount produced, sales, salary level, or level reached in the organization. The crucial distinction between performance and effectiveness is that the latter does not refer to behavior directly but rather is a function of additional factors not under the control of the individual (e.g., state of the economy, nepotism, quality of raw materials, etc.).

It is our contention that psychologists should be trying to measure and predict the

major dimensions of performance rather than effectiveness, since a measure of effectiveness is one or more steps removed from what the individual actually does. A procedure that is directly in line with this objective is the method of scaled expectations first proposed by Smith and Kendall (1963) and since used by Folgi, Hulin, and Blood (1971), Landy and Guion (1970), and Zedeck and Baker (1971), among others. The procedure is a variant of critical incident methodology that requires the appropriate organizational personnel to consider in detail the components of performance for the job in question and to define anchors for the performance continua in specific behavioral terms. Two additional virtues are that the rating scales are developed through extensive participation by the people who will use them, and the resulting language is that of the organization.

The intent of the present study was to develop and evaluate behaviorally based rating scales for the major functions comprising the job of department manager in a retail store and argue for their utility as criteria for selection research, performance factors for appraisal, and definitions of training needs. Specifically, we wanted to determine if such scales would yield less leniency

<sup>1</sup> Requests for reprints should be sent to John P. Campbell, Department of Psychology, University of Minnesota, Elliott Hall, Minneapolis, Minnesota 55455.

<sup>2</sup> Now at the University of Tennessee.

and halo errors than a rating procedure that was not behaviorally anchored and whether they would exhibit significant convergent and discriminant validity. The firm under consideration is a large, nationwide retail chain with the department manager constituting the first level of management. He has responsibility for ordering merchandise, supervising sales personnel, maintaining inventory, and the like.

## METHOD

### *Scale Development*

An initial series of workshops was held with 20 store managers and assistant store managers in the Twin Cities area. The workshop participants were the immediate supervisors of the department managers for whom the rating scales were to be developed. The original Smith and Kendall procedure calls first for the development group to name and define the major components of performance for the job in question. Using these definitions as guides, the participants then are asked to describe specific behavioral episodes that illustrate both effective and ineffective performance (i.e., critical incidents) within each of the a priori factors.

The present study modified this procedure a bit. After a general discussion of problems inherent in performance rating and a description of critical incident methodology, the participants in the first workshop session were asked to write at least five effective and five ineffective critical incidents of department manager performance, with no prior discussion of the underlying performance factors. A pilot workshop suggested that this modification was more effective in keeping the conversation away from a discussion of traits and centered on behavior than it was at the beginning with an attempt to define the major performance factors and then writing incidents to illustrate these factors.

The behavioral incidents produced in the first session were then submitted to a qualitative cluster analysis. That is, the first and second authors sorted the incidents into what appeared to be homogeneous categories and wrote a tentative definition for each category. These tentative definitions of performance dimensions were fed back to the store managers and assistant store managers in a second workshop session. The ensuing discussion centered around (a) whether the tentative factors were meaningful and important, (b) whether there was too much overlap, (c) whether important components of performance were not represented, and (d) whether the definitional language was organizationally correct. As a result, two dimensions were dropped, one was added, and several other dimensions were readjusted, yielding 10 dimensions. Next, the participants were asked to write more behavioral incidents to fill in gaps that appeared to describe episodes representing more moderate levels of performance rather than only extremely effective or ineffective samples.

That part of the above procedure commencing with the second workshop session was repeated with a similar group of 18 store managers and assistant store managers in St. Louis. More incidents were written and the definitions of the dimensions were altered somewhat, although no extensive changes were made.

After the behavioral incidents were edited to remove redundancy and were shortened as much as possible, the retranslation step of the Smith and Kendall procedure was carried out. Each participant in the Minneapolis-St. Paul and St. Louis workshops was asked to make two judgments concerning each incident. First, the definitions of the 10 performance dimensions were presented, and the participants were asked to sort each incident into the dimension that it most closely represented. Second, each incident was rated on a 9-point scale based on the degree of effective or ineffective performances that it represented relative to the performance dimension in which it was grouped. Incidents were retained as defining anchors for the performance dimension if at least 30 of the 38 judges agreed on their classification and if the *SD* of the scale values assigned was less than 1.75. Approximately 30% of the incidents were eliminated by these criteria. One of the 10 original performance dimensions could not be clearly retranslated, and it was subsequently dropped from further consideration. A highly abbreviated definition for each of the nine dimensions surviving the retranslation step is given below:

1. Supervising sales personnel—gives sales personnel a clear idea of their job duties and responsibilities; exercises tact and consideration in working with subordinates, handles work scheduling efficiently and equitably, and supplements formal training with his own "coaching"; keeps himself informed of what his sales people are doing on the job and follows company policy in his agreements with subordinates.

2. Handling customer complaints and making adjustments—informs customers accurately and tactfully of company policy; smooths things over with customers who have complaints such that they will continue to purchase or increase their purchases from the company; sets good examples for sales personnel to follow.

3. Meeting day-to-day deadlines—meets deadlines according to systems developed by higher management; orders merchandise on time to assure proper stock position and gets work schedules for sales personnel planned and recorded on time; plans special promotions so that they get underway according to deadlines.

4. Merchandise ordering—maintains a good mix of colors, styles, and sizes, etc.; develops procedures for keeping track of the merchandise flow; makes advantageous use of company guidelines in ordering decisions and modifies guidelines according to seasonal trends, merchandise flow, and stock position.

5. Developing and planning special promotions—plans promotions carefully, far enough ahead, and in sufficient detail so that he does not overlook important aspects; develops new ideas and new approaches in planning displays and merchandise lay-

	9	Could be expected to conduct a full day's sales clinic with two new sales personnel and thereby develop them into top sales people in the department.
Could be expected to give his sales personnel confidence and a strong sense of responsibility by delegating many important jobs to them.	8	
	7	Could be expected <i>never</i> to fail to conduct training meetings with his people weekly at a scheduled hour and to convey to them exactly what he expects.
Could be expected to exhibit courtesy and respect toward his sales personnel.	6	
	5	Could be expected to remind sales personnel to wait on customers instead of conversing with each other.
Could be expected to be rather critical of store standards in front of his own people, thereby risking their developing poor attitudes.	4	
	3	Could be expected to tell an individual to come in anyway even though she/he called in to say she/he was ill.
Could be expected to go back on a promise to an individual whom he had told could transfer back into previous department if she/he didn't like the new one.	2	
	1	Could be expected to make promises to an individual about her/his salary being based on department sales even when he knew such a practice was against company policy.

FIG. 1. Scaled expectations rating scale for the effectiveness with which the department manager supervises his sales personnel.

outs; plans special promotions and uses them to advantage in selling "old" merchandise.

6. Assessing sales trends and acting to maintain merchandising position—reevaluates sales trends and takes them into account in maintaining an up-to-date merchandising position; takes quick action to gather more information in response to customer requests for new or different items; shops competitors' stores, when appropriate, to gather information about sales trends and customer preferences.

7. Using company systems and following through on administrative operations—makes effective use of company systems and procedures; handles necessary paper work quickly and accurately; knows company, store, and department goals; follows through on invoice files to note that needed items have been

received and follows through on shipments shown to be short or inaccurate.

8. Communicating relevant information to associates and to higher management—keeps store management informed of how things are going and provides information necessary for planning store-wide programs; keeps his sales personnel informed of what's going on in the store; consults with sales personnel about department operations.

9. Diagnosing and alleviating special department problems—quickly recognizes instances of something wrong in the department; goes somewhat beyond the call of duty in sizing up how a department is doing; develops solutions to problems that are innovative and that go beyond prescribed or standardized company or store procedures.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR EACH RATER USING EACH METHOD AND THE CORRELATIONS BETWEEN RATERS WITHIN METHODS

Performance factor	Summated score <sup>a</sup>					Scale score				
	Manager		Assistant manager			Manager		Assistant manager		
	M	SD	M	SD	r	M	SD	M	SD	r
A	3.14	.42	3.10	.46	.58	5.67	.87	5.69	.89	.43
B	3.50	.44	3.52	.45	.42	6.27	1.08	6.14	.99	.31
C	3.21	.48	3.10	.54	.59	6.06	1.01	5.95	1.09	.43
D	3.15	.47	3.10	.50	.54	6.45	1.31	6.23	1.31	.41
E	2.86	.51	2.78	.54	.49	5.52	1.62	5.40	1.56	.48
F	2.89	.55	2.70	.58	.54	6.05	1.19	5.91	1.30	.48
G	3.15	.49	3.06	.54	.57	6.34	1.19	6.17	1.35	.55
H	3.09	.51	2.98	.54	.49	5.93	1.17	5.79	1.29	.39
I	2.68	.58	2.60	.63	.56	5.53	1.24	5.32	1.28	.40

Note. The summated scores are on a 4-point scale, and the scale scores are on a 9-point scale.

<sup>a</sup> Because different numbers of items were used to define the dimensions for the summated scores, the means given here have been converted to the mean response per item.

The finished rating scales for each of the nine dimensions consisted of the scale definition and a 9-point continuum defined by specific behavioral incidents with the appropriate scale values. The scale for supervision is shown in Figure 1. To help obviate the domain sampling problem, each illustrative incident was stated in the form "could be expected to . . .," rather than implying that the person to be rated actually had to exhibit that specific behavior.

Yet a third workshop was held with a group of store managers and assistant store managers in the Chicago area for purposes of reviewing the finished product. Only minor changes in language and definitions were made.

### Alternative Rating Method

A second method for assessing performance on each dimension was developed by using the scale definitions produced in the workshops to construct summated rating scales for each dimension. That is, the definitions produced by the above procedure were broken down into their major elements, and each of these separate statements was used as a Likert-type item with a 4-point response format. All the items were statements or functions that contributed to high performance on a particular dimension, and the individual was rated as exhibiting it very rarely (1) to almost always (4). The number of items varied from 5 to 11, depending on the number of elements the workshop participants had included in the definition of each performance dimension. An individual's rating for a dimension was simply the average item response for that dimension.

For comparative purposes, it should be noted that both the performance dimensions to be rated and their definitions were identical for the two methods. They were what survived the retranslation procedure and the scrutiny of the multiple workshop par-

ticipants. In this sense, both methods dealt with performance rather than effectiveness (in the sense that these two terms were used above). The major difference between them is that the scaled expectations method uses scaled behavioral anchors, and the summated ratings method does not. Since the dimensions used by each method were both products of the same attempt to define performance as carefully as possible, the chances of finding major differences in halo, leniency, etc., should be considerably less than if a more haphazard method had been used for comparative purposes.

### Subjects for Scale Evaluation

The subjects consisted of 537 department managers selected haphazardly from throughout the United States, with the exception of the southeastern region. They varied in age and experience, but the age distribution had a pronounced positive skew.

### Procedure for Scale Evaluation

Each department manager was rated by both his store manager and assistant store manager using both the scaled expectations method and the summated ratings procedure. Each rater was asked to use the scaled expectations procedure to rate both the "typical" and "best" performance of each subject. The purpose of the best rating was to reduce leniency error in the typical rating. In sum, each subject was rated on nine performance dimensions by two raters using two different rating methods.

### RESULTS

The means and standard deviations of the four sets of ratings (2 raters  $\times$  2 methods) are given in Table 1. Also shown are the cor-

TABLE 2

FACTOR MATRIX GENERATED FROM INTERCORRELATIONS BASED ON SUMMATED RATINGS BY STORE MANAGERS

Performance dimension	Squared factor loadings					Four others	$h^2$
	I	II	III	IV	V		
A	.14	.19	.05	.02	.40	.03	.83
B	.04	.28	.02	.01	.02	.02	.39
C	.17	.16	.34	.02	.06	.02	.77
D	.27	.04	.12	.04	.14	.11	.74
E	.31	.09	.07	.02	.01	.23	.73
F	.38	.06	.07	.01	.11	.12	.75
G	.25	.10	.12	.19	.01	.14	.81
H	.23	.21	.06	.02	.01	.23	.76
I	.50	.10	.04	.01	.00	.06	.71

Note. We show squared factor loadings in Tables 2 and 3 because we wish to illustrate the relative contribution of each performance dimension to total variance in each of the factors. In addition, the squared loadings show more clearly than loadings how the communality for each of the dimensions is arrayed across the various factors.

The highest squared loading in each row is circled with the exceptions of scales D and F, which are loaded equally on two factors.

relations between raters for each performance dimension within each rating method.

In general, the leniency error was not severe for the method of scaled expectations but was rather pronounced for several factors rated via summated ratings. The maximum possible summated ratings score is 4.0, and six of the nine scales yielded means between 3.0 and 4.0. The customer relations scale (B) was the worst offender. In contrast, the mean ratings using the scaled expectations method clustered around 6.0, which is reasonably close to the midpoint of 5.0. Relative to differences in raters, store managers tended to give slightly higher ratings than assistant store managers, regardless of the method.

The correlations across raters within method were not high, and they were somewhat lower for the scaled expectations method. Given the assumption that each rater possesses similar knowledge about each ratee, this correlation could be viewed as an index of interrater agreement. However, as will be pointed out below, such an assumption may not be warranted. The interpretation of the difference between the correlations for the two methods is also confounded by the fact that there may be more method variance

incorporated in the summated ratings than in the scaled expectations.

A major line of support for the scaled expectations technique is derived from factor analyses of the four sets of ratings. Four  $9 \times 9$  correlations matrices were generated and factor-analyzed via the principal factors technique with squared multiple correlations as communality estimates and with the stipulation that nine factors must be extracted, regardless of the level of common variance. Each solution was rotated to simple structure via the varimax procedure.

The matrices of squared factor loadings for the store managers' ratings using both the summated rating and scaled expectations techniques are shown in Tables 2 and 3. The clearer solution was obtained from the store manager ratings using the scaled expectations technique. That is, the procedure tended to yield nine nontrivial factors with one high loading per factor. However, factor VI is very weakly defined with a loading of only .30 on performance dimension E.

The solution obtained from the summated ratings yielded a much larger general factor that could not be broken up by forcing the common variance into nine factors. A similar

TABLE 3

FACTOR MATRIX GENERATED FROM INTERCORRELATIONS BASED ON  
SCALED EXPECTATION RATINGS BY STORE MANAGERS

Performance dimension	Squared Factor Loadings									$h^2$
	I	II	III	IV	V	VI	VII	VIII	IX	
A	.04	.05	.01	.01	.04	.00	.01	.01	.36	.54
B	.02	.35	.01	.01	.02	.00	.00	.00	.03	.45
C	.07	.04	.02	.01	.27	.00	.01	.01	.08	.53
D	.10	.06	.02	.15	.05	.01	.01	.01	.15	.57
E	.10	.07	.04	.04	.08	.09	.01	.01	.12	.57
F	.10	.06	.04	.01	.17	.00	.02	.17	.10	.69
G	.40	.05	.03	.02	.04	.00	.01	.01	.07	.66
H	.10	.08	.04	.02	.10	.01	.16	.03	.13	.69
I	.15	.09	.23	.01	.07	.01	.01	.02	.04	.64

Note. The highest squared loading in each row is circled with the exceptions of scales D and F, which are loaded equally on two factors.

general factor was found when ratings of the assistant store managers were analyzed, regardless of method.

The final mode of analysis was the multitrait, multimethod approach suggested by Campbell and Fiske (1959). The present study produced a  $36 \times 36$  multitrait (9 performance dimensions), multimethod (summed ratings vs. scaled expectations), multirater (store managers vs. assistant store managers) matrix. Since the factor analyses indicated that the ratings by the managers yielded somewhat clearer factor structure than the assistant managers' ratings, only the  $18 \times 18$  multitrait, multimethod matrix for the managers' ratings is shown in Table 4.

In terms of convergent and discriminant validity, Table 4 indicates that significant convergent validity was achieved. Campbell and Fiske define convergent validity as *the observation of significant correlations when two different methods are used to measure the same variable*. All the entries in the validity diagonal are significantly different from zero at alpha = .001. Discriminant validity is indicated in two ways. First, the entries in the validity diagonal can be compared with

their corresponding row and column entries in the heterotrait-heteromethod triangles. This yields 18 comparisons for each performance factor in which the diagonal value should be higher than the row and column values, if discriminant validity is present. The diagonal entry is higher for 136 out of 144 such comparisons. Scale D (merchandise ordering) accounts for six of the eight discrepancies.

A second index of discriminant validity involves comparing the validity diagonal entries (same trait but different methods) to the corresponding row and column entries in the heterotrait-monomeethod triangles. This implies that the correlations should be higher when different methods are used to measure the same dimension than when different dimensions are measured by the same method. For the summed ratings method, only 16 of the 72 comparisons yielded higher entries in the validity diagonal. In contrast, the validity entry was higher in 60 of 72 comparisons for the scaled expectations method. Of the 12 discrepancies for scaled expectation, 6 were due to scale D and 4 to scale F.

Similar levels of discriminant validity were

TABLE 4

MULTITRAIT (PERFORMANCE DIMENSIONS), MULTIMETHOD (SUMMATED RATINGS VERSUS SCALED EXPECTATIONS) MATRIX FOR STORE MANAGER RATINGS

Method

Method

Summated ratings

Scaled expectations

Method	Summated ratings								Scaled expectations									
	A	B	C	D	E	F	G	H	I	A	B	C	D	E	F	G	H	I
A																		
B	.55																	
C	.72	.42																
D	.64	.41	.71															
E	.67	.42	.67	.66														
F	.64	.39	.61	.70	.70													
G	.69	.42	.76	.73	.72	.70												
H	.75	.46	.66	.65	.70	.67	.70											
I	.65	.35	.58	.66	.68	.70	.66	.67										
<hr/>																		
Summated ratings																		
<hr/>																		
Scaled expectations																		
<hr/>																		
Multidimensional-Heterodimensional-Heteromethod Triangle, and																		
<hr/>																		
Multidimensional-Heterodimensional-Monomethod Triangle.																		
<hr/>																		

Note: N = 527; O = Validity Diagonal; L = Heterodimensional-Heteromethod Triangle, and

found when the rating method is held constant and the multitrait, multirater matrices are examined.

### DISCUSSION AND CONCLUSIONS

It seems fair to conclude from the above data that this variant of the scaled expectations procedure produced performance ratings that were not subject to many of the errors commonly associated with such ratings. There was less leniency error, less halo error, and less method variance than that produced by the summated ratings method. The ability of the scaled expectations method to produce the factor structure shown in Table 3 is gratifying. However, both the factor analysis and the multitrait, multimethod matrix indicate that scales D and F are less satisfactory than the others, in terms of method variance and discriminant validity.

One possible explanation for the lack of clarity in the assistant store managers' ratings is that his job duties are more heavily loaded toward the merchandising function rather than supervision. The store manager has direct responsibility for the supervision of department managers and, other things being equal, he should be more familiar with their performance. This would tend to explain the relatively low correlations between store managers and assistant store managers and preclude their being used as indices of interrater agreement.

Several outcomes that are not reflected in the empirical results also deserve mention. The managers who developed these scales invested a tremendous amount of effort in the process, and it seemed to be a valuable learning experience for them. It is our contention that most people in organizations seldom, if ever, give careful attention to what they really mean by effective performance. The above procedure forces a confrontation with this question. Defining effective performance and assessing subordinates accordingly is an integral part of management. The workshop participants realized this early in the proceedings and did not try to circumvent or otherwise avoid this difficult task. We believe that the above procedure is an effective vehicle for facilitating this confrontation and for devel-

oping an appreciation for the need to talk about performance in behavioral terms.

Potentially, there are many applied uses for the outputs of this procedure. The scales can serve as criteria against which to evaluate predictors for selection and promotion decisions. They could also profitably be incorporated in performance appraisal and review systems. By virtue of the way they were developed, they represent behavioral specifications of desired behavior and, at the same time, provide a metric for person-to-person comparisons. Thus, they avoid some of the problems of both the traditional kind of performance appraisal and mutual goal setting. As pointed out by McGregor (1957), traditional performance appraisal suffers from a lack of behavioral specifications that makes it very difficult to give feedback to individuals or plan how their performance could be improved. Mutual goal setting helps solve this problem but makes person-to-person comparisons very difficult. Finally, the efforts of the workshop participants could also be viewed as defining desired behaviors around which a training and development system could be organized.

### REFERENCES

- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.
- FOLGI, L., HULIN, C. L., & BLOOD, M. R. Development of first-level behavioral job criteria. *Psychological Bulletin*, 1971, 55, 3-8.
- LANDY, F. J., & GUION, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, 5, 93-103.
- MCGREGOR, D. An uneasy look at performance appraisal. *Harvard Business Review*, 1957, 35, 89-94.
- SMITH, P., & KENDALL, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, 47, 149-155.
- ZEDECK, S., & BAKER, H. T. Evaluation of behavioral expectation scales. Paper presented at the meeting of the Midwestern Psychological Association, Detroit, May 1971.

(Received July 19, 1971)

## ESTIMATING THE INFLUENCE OF JOB INFORMATION ON INTERVIEWER AGREEMENT

JOHN A. LANGDALE AND JOSEPH WEITZ<sup>1</sup>

*New York University*

Two groups of personnel interviewers were given eight application blanks to judge. One group was given a general job title only, the other group a rather full description of the job to be filled. The interrater reliability was far superior for the group having more complete job information; there was also greater discrimination among applicants for this group. Length of service of the interviewers had little effect.

The primary purpose of this study is to examine the influence of job information on personnel selection decisions. Do personnel selectors who are given exact information about the job actually manifest more agreement among themselves in rating overall suitability of a candidate than do those selectors who are given merely a general job title?

As an exploratory study, application blanks were evaluated by the interviewers in this study since there are fewer uncontrollable variables than in a face-to-face interview. This is a similar strategy to those that resort to written descriptions of candidates (Carlson, 1967; Carlson & Mayfield, 1967; Mayfield & Carlson, 1966; Miller & Rowe, 1967; Rowe, 1963, 1967), protocols (Bolster & Springbett, 1961), and resumés (Hakel, Dobmeyer, & Dunnette, 1970; Hakel, Ohnesorge, & Dunnette, 1970). This group of related studies uses procedures somewhat similar in nature to those used in this study. Where possible, an attempt has been made to consider the following variables because those studies suggest them as systematic sources of variance in interviewers' decisions: the order of candidate presentation, order or primacy effects of individual items of information, the range of candidate sample, relative quota situation, interviewer leniency as a trait, interviewer experience, interrater-intrarater agreement on individual items, content dimensions of items, and importance or weight of the items of information concerning the applicant.

Particularly, in addition to determining the effect of job information on interrater agreement, an analysis of the content of the application blank will be made in terms of what items interviewers feel most important and whether such appraisals change as a result of different amounts of information.

### METHOD

#### Subjects

Sixty-two interviewers from various public and private employment facilities in New York City were asked to participate. Out of this group, 33 accepted although later in the project 3 subjects had to be dropped since instructions were not followed exactly. All subjects were female to avoid any systematic sex differences over the experimental treatments. This group of 30 subjects was randomly split: 15 received exact job information (median age = 35, median years of education = 16.2, median years of experience = 3.5) and 15 received only a job title (median age = 29, median years of education = 16.9, median years of experience = 3).

#### Materials Used

An application blank consisting of 18 discrete items of information was constructed attempting to preserve the content and format of a prototypic application blank. Two different scales were devised to measure (a) the estimated importance of each item response to overall evaluation of the candidate's application (a unipolar, 4-point scale from "extremely important" to "neutral") and (b) a final rating scale of overall suitability of candidates for the job available (a bipolar, 7-point scale from "extremely qualified" to "extremely unqualified"). Under each scale point was printed a verbal description, exemplifying various degrees of each dimension as an anchor.

Eight hypothetical applicants were constructed by having eight secretaries fill in copies of the application. These eight application blanks, each representing an applicant, served as the stimulus materials for the

<sup>1</sup> Requests for reprints should be sent to Joseph Weitz, Department of Psychology, New York University, 21 Washington Place, Room 300, New York, New York 10003.

subjects. Since the same eight applicants would be rated under two different conditions, they were chosen to provide a heterogeneous candidate range for both conditions. But such factors as item favorability, item importance, and interrater agreement on them was not "rigged" or preestablished as has been the case in some studies (Carlson, 1967; Carlson & Mayfield, 1967; Mayfield & Carlson, 1966; Rowe, 1963). This was avoided so as to gain a more natural index of agreement among subjects.

From the above materials, a test booklet was constructed containing a set of detailed instructions, a page consisting of pertinent background questions about subjects, and eight scales on which to rate overall suitability of each candidate. The remaining pages were made up of the eight applications and attached scales to rate importance of the 18 items within each blank.

### Procedure

All 30 subjects, "tested" individually, were given the same eight applicants to evaluate. In the written instructions, all subjects were asked to consider each application independently by reading it once, then rating item importance, and, finally, assessing that candidate's overall qualifications. To standardize the relative quota situation, all subjects were told that only one position was open. However, to ensure careful evaluation of each blank, subjects were advised of the equal importance of all judgments. To this extent, all subjects were treated in the same way.

To examine the effects of specificity of job information, 15 subjects were given instructions that read only, "The eight applicants here represented by their application blanks are applying for the position of Secretary." In contrast, the other 15 subjects were given much more explicit information: "The eight applicants . . . are applying for the position of Executive Secretary. The requirements are typing speed of 60 wpm, stenography speed of 100 wpm, dictaphone use, and bilingual ability in either French, German, or Spanish . . . salary: \$10,000 per year." Order or primacy effects of individual items were held constant since all subjects were given the same eight blanks. The order of candidate presentation (as they appeared in the booklet) was randomized,

but each group of subjects received the same randomization, producing "yoked" groups to control for any disproportionate error due to order over the two treatments.

### RESULTS

Overall suitability scores fall into a  $2 \times 8$  factorial with the eight repeated ratings made by each of 30 subjects. The results of this analysis are found in Table 1 along with associated eta-square values indicating the proportion of variance in the ratings accounted for by the independent variables (Cohen, 1965). The effect of the two levels of specificity in job information, our main interest, is significant and accounts for 53% of the sample variance, showing that the information given the interviewer very much influences his assessments of candidates. The interaction, significant and responsible for 18% of the variance, informs us that among the eight applicants there is no consistent pattern of differences between the two groups' evaluations. Generally, being deprived of information, interviewers tend to give higher ratings with less discrimination among candidates; on certain candidates, however, there are disagreements between the two groups that do not conform to this overall pattern.

For the clearest index of interrater agreement in each experimental condition, the previous factorial was merely broken in half, allowing for separate analyses of the two groups and computation of two intraclass correlations (Guilford, 1954). Table 2 presents the results of this procedure. Because of the extreme statistical significance of these findings, we can conclude that interviewers who have been provided with job details display greater interrater agreement ( $r = .87$ ) than do those given only a job title ( $r = .35$ ), even though both groups show significant agreement.

Because of the inconclusive evidence concerning the effect of interviewing experience on interrater agreement using similar materials, an informal analysis on this basis was performed. The two experimental groups were each subdivided into the five most experienced and the five least experienced subjects. Already knowing that more informed selectors show greater interrater agreement and sup-

TABLE 1  
ANALYSIS OF VARIANCE OF OVERALL  
SUITABILITY ASSESSMENTS

Source	df	MS	F	eta
Between groups	1	130.54	31.85*	.53*
Error	28	4.098		
Between candidates	7	53.895	63.41*	.57*
Candidates $\times$ groups	7	16.004	18.83*	.18*
Error	196	0.8499		

\*  $p = .01$ .

posing that experience increases a judge's reliability, one might predict the following values for the groups' intraclass correlations: informed-experienced > informed-inexperienced > uninformed-experienced > uninformed-inexperienced. However, the results appear to contradict expectations since the correlations of the four groups are respectively .81, .92, .18, and .56.

To gain some insight into the internal structure of the application blank, the last analysis investigates the 18 items within the blank that our interviewers considered most important to their final judgments of the candidates. Given our two information conditions and a mean item importance rating computed across the eight blanks for each subject, the resulting 540 means were put into a  $2 \times 18$  factorial, yielding the contents of Table 3. The significance between groups difference indicates that the estimated importance of items is a function of specificity of job information. However, the distinction between items accounts for a larger proportion (45%) of sample variance. Less informed interviewers generally rate items as less important; nevertheless, there is a marked tendency, regardless of knowledgeability to rate the same items as either of high or low importance, which explains the very significant between item *F* ratio and the small amount of variance accounted for by the interaction effect.

Items found to be most important by both groups included the type of position sought

TABLE 3  
ANALYSIS OF VARIANCE OF IMPORTANCE  
OF ITEMS SCORES

Source	df	MS	F	eta <sup>2</sup>
Between groups	1	22.72	12.78*	.31*
Error	28	1.78		
Between items	17	9.85	24.38	.45*
Items $\times$ groups	17	0.933	2.31*	.04*
Error	476	0.404		

\*  $p = .01$ .

and salary expected by the applicant, whether she desired full-time, permanent employment, what secretarial skills she had such as typing and stenography speed, and the place of last employment (here an interesting trend formed in weighting progressively less, the further the position was held in the past). Large differences between the groups occurred in the perceived importance of physical attributes, marital status, number of children, and general outside interests, but each of these items was given more weight by subjects with exact job information.

## DISCUSSION

Holding constant the possible order effects of component items of information, their content dimensions, the range of candidate sample, the relative quota situation, and matching the random order of candidate presentation, we have found the hypothesis of major concern to be clearly substantiated by the results. Personnel interviewers furnished with more exact job information showed a much higher degree of interrater reliability on overall applicant assessments. That reliability ( $r = .87$ ) far exceeded expectations; although Carlson (1967), "rigging" the items in a written description so that all judges agreed on their favorability, found a coefficient as high as .90; Hakel, Dobmeyer, and Dunnette (1970), employing methods and materials like our own, found a less artificial reliability coefficient of .68. The practical implications of the results here seem clear—by availing the interviewer of rather extensive information about the job to be filled, such as that provided by detailed job descriptions and job

TABLE 2  
ANALYSIS OF THE TWO GROUPS INDEPENDENTLY

Source	df	MS	F	n
Interviewers given exact job description				
Between subjects	14	2.96	5.0*	
Between candidates	7	60.03	101.41*	.87*
Error	98	.592		
Interviewers given nonspecific job description				
Between subjects	14	5.24	4.73*	
Between candidates	7	9.87	8.91*	.35*
Error	98	1.11		

\*  $p = .01$ .

tities, reliability of employment selection decisions can be increased.

Equally as evident are the consequences of depriving personnel interviewers of details about the job they are screening for—a lack of discrimination among applicants ensued, and, at least for certain candidates, evaluations were even inconsistent with those made under more informed conditions. Also, from the analysis of item importance, interviewers thus deprived tended to assign less importance to candidates' responses, although overall appraisals were typically more lenient under this condition.

Less satisfactory are the results on interviewer experience and its effects on the reliability of candidate appraisals. Using resumés, Hakel et al. (1970) found a higher intraclass correlation (.68) for interviewers than for students (.48) examining the same materials. Comparing life insurance managers on the basis of type and length of experience, Carlson (1967) found no difference in either intra- or interrater agreement on written descriptions. The informal analysis performed here on a limited sample contradicted the post hoc hypothesis, revealing somewhat less reliability among veteran interviewers. Of course, even granting that they are not in agreement on most candidates, this does not preclude the possibility that experienced interviewers may be validly selecting those applicants who would later have more success on the job. These issues certainly deserve more thorough treatment, but, at present, the only conclusion to be drawn is that experience, per se, does not appear to be a strong predictor of reliability in candidate assessment.

The application blank itself, as a selection device, exhibits certain noteworthy qualities beyond its ability to generate superior degrees of interrater reliability than typically associated with face-to-face interviews. Despite the specificity of job information, certain items within the blank characteristically receive more subjective weight from judges. This would indirectly support Hakel et al. (1970) in their contention that, regardless of the many sources of systematic influence on overall judgments, there seem to be certain kinds

of information consistently made use of by interviewers. Looking to the type of items stressed by our judges and those in the study just mentioned, the weight assigned seems, in most cases, to be a function of the particular job the interviewers think they are screening for.

Throughout this discussion, however, only the reliability side of the coin has been showing—the validity of overall assessments as a function of available information is just as important, if not more so. Another limitation is our exclusive use of female subjects; thus, the question as to whether male judges would manifest the same behavior must be left open. Finally, in our attempt to merely estimate the influences of job information on interviewer agreement, we have failed to isolate the separate contributions of job title and job description, both of which were manipulated as components affecting judges' knowledge of the job. Further, since this was an exploratory study, only the extremes of information were used in order to see if any effect existed as a result of the manipulation. Obviously there was an effect, and it is felt that various degrees of job information will have a similar effect in the interview situation.

#### REFERENCES

- BOLSTER, B. I., & SPRINGBETT, B. M. The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology*, 1961, 45, 97-103.
- CARLSON, R. E. Selection interview decisions: The effect of interviewer experience, relative quota situation, and applicant sample on interviewer decisions. *Personnel Psychology*, 1967, 20, 259-280.
- CARLSON, R. E., & MAYFIELD, E. C. Evaluating interview and employment application data. *Personnel Psychology*, 1967, 20, 441-460.
- COHEN, J. Some statistical issues in psychological research. In B. Wolman (Ed.), *Handbook of clinical psychology*. New York: McGraw-Hill, 1965.
- GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.
- HAKEL, M. D., DOBMEYER, T. W., & DUNNETTE, M. D. Relative importance of three content dimensions in overall suitability ratings of job applicants' resumes. *Journal of Applied Psychology*, 1970, 54, 65-71.

- HAKEL, M. D., OHNESORGE, J. P., & DUNNETTE, M. D. Interviewer evaluations of job applicants' resumes as a function of the qualifications of the immediately preceding applicants: An examination of contrast effects. *Journal of Applied Psychology*, 1970, 54, 27-30.
- MAYFIELD, E. C., & CARLSON, R. E. Selection interview decisions: First results from a long-term research project. *Personnel Psychology*, 1966, 19, 41-53.
- MILLER, J. W., & ROWE, P. M. Influence of favorable and unfavorable information upon assessment decisions. *Journal of Applied Psychology*, 1967, 51, 432-435.
- ROWE, P. M. Individual differences in selection decisions. *Journal of Applied Psychology*, 1963, 47, 304-307.
- ROWE, P. M. Order effects in assessment decisions. *Journal of Applied Psychology*, 1967, 51, 170-173.
- WEBSTER, E. C. Decision making in the employment interview. *Personnel Administration*, 1959, 22, 15-22.

(Received August 23, 1971)

# LEADER BEHAVIOR MEASUREMENT IN GERMAN INDUSTRY

D. TSCHEULIN<sup>1</sup>

*University of Würzburg, Federal Republic of Germany*

The two main factors of Consideration and Initiating Structure as they relate to supervisory behavior appear to be confirmed as important dimensions in West Germany, as has been indicated in other European Countries (e.g., Sweden (Lennerlöf, 1965) and the Netherlands (Philipsen, 1965)). Some questions remain about other and more specific dimensions and their possible relations to other organizational variables (e.g., level in the administrative hierarchy). Studies on these questions are continuing in Germany.

The Supervisory Behavior Description (SBD) Questionnaire was first described by Fleishman (1951, 1953, 1957) and has received widespread use in many industrial settings in the United States and abroad. The questionnaire measures the now well known dimensions of "Consideration" and "Initiating Structure" identified in the Ohio State Leadership studies (Stogdill & Coons, 1957). Recent definitions of these dimensions are in Fleishman and Simmons (1970).

In recent years, there has been increased interest in the application of these concepts and measures to industrial supervisors and managers in Germany. The purpose of this article is to present some recent research that examines the applicability of the concepts of Consideration and Initiating Structure to the description of supervisory behavior in German industry.

Specifically, we first report a replication of Fleishman's factor analysis of supervisory behavior description items and, second, review some related research with this questionnaire in German industrial situations.

## METHOD

A German translation by Tscheulin and Rausche (1970) of Fleishman's (1953, 1957) SBD Questionnaire was administered to 183 employees who described 44 immediate supervisors from two different firms located in West Germany. Seventy-four of the subjects were employed by a publications firm and 102 by a marketing research institute.

The German version of the SBD Questionnaire was adapted with the intention of not only achieving a translation that would be linguistically correct but

<sup>1</sup> Requests for reprints should be sent to D. Tscheulin, Department of Psychology, University of Würzburg, Hofstrasse 10, 87 Würzburg, Federal Republic of Germany.

also one that would retain the psychological implications (intentions) of the individual items. The format and scoring procedure of the original questionnaire were preserved.<sup>2</sup>

## RESULTS

The correlations among the 48 items, obtained from the 183 questionnaire responses, were subjected to a principal-axis factor analysis. Subsequent orthogonal rotations using the varimax method indicated that a two-factor solution could be accepted. (The eigenvalues, in order of factor extraction, were 9.6, 7.5, 1.7, 1.6 . . .). Table 1 presents the orthogonal loadings of each item of each of the two factors. Also presented are the loadings originally obtained by Fleishman (1951, 1953) with these same items in the original standardization of the questionnaire.

As can be seen in Table 1, the factor loadings on each factor correspond well with the American findings. The factorial similarity between the German and the American samples were tested by the phi-coefficient following the Wrigley and Neuhaus procedure described in Harman (1967, p. 270) and found to be  $\phi = 0.91$  for Consideration and  $\phi = 0.81$  for Initiating Structure. It is also noted that out of the 48 original items only a few would not have been selected in the same scale from the German analysis.

Although reliabilities could probably be improved through further item analysis, split-

<sup>2</sup> In contrast to the U. S. questionnaire, where each item was responded to on a 5-point scale (e.g., always, often, occasionally, seldom, never), the German version used a 6-point scale. As in the U. S. version, scaling procedures were utilized with frequency adjectives to achieve "equal appearing" intervals between German frequency adverbs.

TABLE 1  
FACTOR LOADINGS OF SUPERVISORY BEHAVIOR DESCRIPTION ITEMS OBTAINED  
FROM AMERICAN AND GERMAN SAMPLES

Consideration items	Sample			
	American*		German	
	Consideration	Initiating Structure	Consideration	Initiating Structure
He refuses to give in when people disagree with him.	-.68	.06	-.38	.23
He does personal favors for the foremen under him.	.40	.06	.65	.26
He expresses appreciation when one of us does a good job.	.70	.19	.55	.12
He is easy to understand.	.70	.13	.63	.01
He demands more than we can do.	-.40	-.08	-.45	.45
He helps his foremen with their personal problems.	.32	.05	.54	.21
He criticizes his foremen in front of others.	-.49	.03	-.60	.34
He stands up for his foremen even though it makes him unpopular.	.54	.08	.71	.19
He insists that everything be done his way.	-.52	-.01	-.38	.42
He sees that a foreman is rewarded for a job well done.	.70	.05	.40	.30
He rejects suggestions for changes.	-.62	-.06	-.54	.22
He changes the duties of people under him without first talking it over with them.	-.69	.09	-.56	.25
He treats people under him without considering their feelings.	-.72	.41	-.70	.34
He tries to keep the foremen under him in good standing with those in higher authority.	.68	.17	.71	.30
He "rides" the foreman who makes a mistake.	-.61	.37	-.51	.43
He refuses to explain his actions.	-.72	.23	-.45	.19
He acts without consulting his foremen first.	-.73	.01	-.67	-.01
He stresses the importance of high morale among those under him.	.73	-.11	.34	.51
He backs up his foremen in their actions.	.62	.16	.64	.18
He is slow to accept new ideas.	-.66	-.06	-.21	.04
He treats all his foremen as his equal.	.66	.28	.69	-.08
He criticizes a specific act rather than a particular individual.	.63	.14	.19	-.07
He is willing to make changes.	.78	.09	.55	-.10
He makes those under him feel at ease when talking with him.	.86	.17	.68	.01
He is friendly and can be easily approached.	.82	-.02	.71	.05
He puts suggestions that are made by foremen under him into operation.	.87	.11	.55	.26
He gets the approval of his foremen on important matters before going ahead.	.65	-.02	.49	.31

*continued p. 30*

TABLE 1 (*Continued*)

Initiating structure items	Sample			
	American*		German	
	Consideration	Initiating Structure	Consideration	Initiating Structure
He encourages overtime work.	.20	.40	-.03	.43
He tries out his new ideas.	-.10	.42	.35	.38
He rules with an iron hand.	-.20	.58	-.37	.58
He criticizes poor work.	-.18	.59	-.18	.63
He talks about how much should be done.	-.20	.60	-.09	.55
He encourages slow-working foremen to greater effort.	.17	.33	.07	.74
He waits for his foremen to push new ideas before he does.	-.07	-.28	.07	.23
He assigns people under him to particular tasks.	.00	.26	.08	.29
He asks for sacrifices from his foremen for the good of the entire department.	.00	.46	.09	.50
He insists that his foremen follow standard ways of doing things in every detail.	.25	.72	-.02	.63
He sees to it that people under him are working up to their limits.	-.17	.87	.06	.73
He offers new approaches to problems.	.36	.72	.49	.30
He insists that he be informed on decisions made by foremen under him.	.13	.51	.11	.57
He lets other do their work the way they think best.	-.17	-.33	.44	-.20
He stresses being ahead of competing work groups.	.03	.34	-.16	.42
He "needles" foremen under him for greater effort.	-.17	.50	-.06	.67
He decides in detail what shall be done and how it shall be done.	.37	.63	-.05	.49
He emphasizes meeting of deadlines.	.10	.68	.04	.49
He asks foremen who have slow groups to get more out of their groups.	-.22	.40	-.03	.70
He emphasizes the quantity of work.	.17	.51	-.21	.69

\* Data from Fleishman (1953).

half reliabilities using all the original items were  $r = .92$  for Consideration and .87 for Structure.

#### DISCUSSION

The similarity of results, across almost 20 years, with different cultures and different methods of analysis, can be considered remarkable. In particular, it may be noted that no maximum approximation of the prior

factor structure, using rotations for maximum congruence, was attempted. Furthermore, Fleishman's original loadings, obtained in 1951, used a Wherry-Gaylor iterative factor analysis solution. The refined methods and computer techniques available today, including better means of communality estimation, produced the present solution. Thus, the two-factor solution appears independent of method.

The findings of the present study are consistent with those of Fleishman concerning the degree of "purity" of the factors. The factorial "similarity" of the Consideration and Initiating Structure factors are very low for both the American sample ( $\phi = 0.07$ ) and the German sample ( $\phi = 0.07$ ). Thus, the independence of these dimensions is demonstrated in both countries.

It may be useful to review some other recent work with this questionnaire in Germany. Using a large sample (1313 subordinates describing 228 supervisors), Fittkau-Garthe (1970) used 38 items, some derived independently and some from the SBD Questionnaire, as a basis for describing supervisors by subordinates. Using different principal-axis factor analysis solutions (2-5 factors) with varimax rotation, the results could be interpreted best in terms of a four-factor solution. All items of the SBD Questionnaire, contained in two of the factors, "Friendly Attention" and "Granting Genuine Participation," were from the original Consideration scale. Similarly, the factors they name as "Work Stimulating Activity" and "Control versus Laissez-Faire" contain items of the original Initiating Structure factor.

Hoefert (1971), investigating relations between supervisory behavior and emotional reactions of subordinates in Germany, also used a larger pool of items and preferred a four factor solution. However, the two main factors that emerged in his work seem comparable to the original Consideration and Initiating Structure dimensions. Lück (1970), using students, found two factors that likewise correspond to the factors Consideration and Initiating Structure. Subsequently, using the German SBD Questionnaire translation carried out by Tscheulin and Rausche (1970), Nachreiner and Lück (1971) conducted item and factor analyses on several samples of students, workers, foremen, general foremen, and managers. Again, they arrived at a two-factor solution, whereby

Factor I could easily be interpreted as Consideration and Factor II as Initiating Structure.

#### REFERENCES

- FITTKAU-GARTHE, H. Die Dimensionen des Vorgesetztenverhaltens und ihre Bedeutung für die emotionalen Einstellungsreaktionen der unterstellten Mitarbeiter. Unpublished dissertation, Universität Hamburg, 1970.
- FLEISHMAN, E. A. Relationship between supervisory behavior and leadership climate. Columbus, Ohio: Ohio State University Personnel Research Board, 1951.
- FLEISHMAN, E. A. The description of supervisory behavior. *Journal of Applied Psychology*, 1953, 37, 1-6.
- FLEISHMAN, E. A. A leader behavior description for industry. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement*. Columbus: Ohio State University, 1957.
- FLEISHMAN, E. A., & SIMMONS, J. Relationship between leadership patterns and effectiveness ratings among Israeli foremen. *Personnel Psychology*, 1970, 23, 169-172.
- HARMAN, H. H. Modern factor analysis. (2nd ed.) Chicago and London: University of Chicago Press, 1967.
- HOEFERT, H. W. Darstellung einer Untersuchung über die Anwendbarkeit von Fragebogen zur Beurteilung des Vorgesetztenverhaltens im Industriebereich. Unpublished paper, Institute für Psychologie (FU), 1971.
- LENNERLÖF, L. Consideration and structure: Two dimensions of supervisory behavior. *Psychological Research Bulletin*, 1965, 5, 1-22.
- LÜCK, H. E. Einige Determinanten und Dimensionen des Führungsverhaltens. *Gruppendynamik*, 1970, 1, 63-69.
- NACHREINER, F., & LÜCK, H. E. Experiences with a German translation of Fleishman's Ohio Leadership Questionnaire. Paper presented at the NATO Symposium on Leadership and Management Appraisal, Brussels, August 2-6, 1971.
- PHILIPSEN, H. Het meten van leiderschap. *Mens en onderneming*, 1965, 19, 153-171.
- STOGDILL, R. M., & COONS, A. E. (Eds.) *Leader behavior: Its description and measurement*. Columbus: Ohio State University, Bureau of Business Research, 1957.
- TSCHEULIN, D., & RAUSCHE, A. Beschreibung und Messung des Führungsverhaltens in der Industrie mit der deutschen Version des Ohio-Fragebogens. *Psychologie und Praxis*, 1970, 14, 49-64.

(Received August 10, 1971)

## EFFECTS OF DIFFERENT LEADERSHIP STYLES ON GROUP ACCURACY<sup>1</sup>

JOSEPH A. CAMPALLERI,<sup>2</sup> HAL W. HENDRICK, WAYNE C. PITTMAN, JR.,  
HARRY D. BLOUT, AND DIRK C. PRATHER

*United States Air Force Academy*

Variables of Fiedler's Contingency Model were manipulated in a group-problem-solving situation. Subjects made private individual estimates of rank-order merit of survival items and subsequently were placed in 48 groups of 4 or 5 to arrive at consensual estimates. Leaders had been contacted earlier, given the solution, and told to assume specific roles: Type I (high accuracy/authoritarian); Type II (high accuracy/democratic); Type III (low accuracy/authoritarian); Type IV (low accuracy/democratic). Type I produced the highest accuracy, Types II and IV had intermediate and comparable accuracy, and Type III produced the lowest accuracy.

Leadership research has shifted from emphasis on personal traits to a conception of leadership as a function of group and environmental variables (Hollander & Julian, 1969). Current avocations such as Olmstead's (1967) leader adaptability are useful. Olmstead contends that specific leadership style is less important than the ability to analyze situational and environmental variables and adapt one's behavior appropriately. Additionally, conceptual frameworks such as Hersey and Blanchard's (1969) Life Cycle Theory, and Fiedler's (1967) Contingency Model have provided heuristic and pragmatic experimental bases for contemporary leadership research. These theories are simultaneously dissimilar and complementary. Based on research, both deny that any one type of leadership style will be universally successful across all personal, group, and environmental variables. Both are experimentally oriented with emphasis on comparisons of effectiveness when contrasting democratic styles with authoritarian styles, thus continuing research initiated by Lewin, Lippit, and White (1939), among others.

Dissimilarities evolve from differences in

<sup>1</sup> The views expressed herein are those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

<sup>2</sup> Requests for reprints should be sent to Joseph A. Cammalleri, Department of Life and Behavioral Sciences, USAF Academy, Department of the Air Force, USAF Academy, Colorado 80840.

basic assumptions of each theory. Life Cycle Theory contends that maturity of the group (psychological age) is the primary determinant of effective leadership style whether democratic (concern for people, relationships) or authoritarian (concern for task, production, autocracy) and that the style is synonymous with *behavior* rather than personality. Consequently, if the leader properly employs diagnostic skills, he may accurately estimate the group's maturity level and employ the appropriate leadership style regardless of his own personality tendencies. Based on a curvilinear progression through four quadrants of the authoritarian and democratic leadership dimensions, the resultant leadership style could then be authoritarian, democratic, or a combination of both.

On the other hand, Fiedler states that the leader's underlying personality structure and tendencies constitute dominant constraints for successful leadership. He advises leaders to seek positions primarily on compatibility of personality with organizational and environmental variables in order to maximize probability of leader success. Thus, Fiedler equates leadership style with *personality* (a predisposition to respond) and not with behavior, as do Hersey and Blanchard. This is an important difference with significant consequences for organizational leadership research.

Fiedler's notion of "favorability" of the environment is critical to his concept of leadership. Specifically, the environment can be ordered categorically according to the degree

Rationale	Solution rank	Survival item
Little or no use on moon	14	Box of matches
Supply daily food required	4	Food concentrate
Useful in tying injured together, help in climbing	6	50 feet of nylon rope
Shelter against sun's rays	8	Parachute silk
Useful only if party landed on dark side	12	Portable heating unit
Food, mixed with water for drinking	11	One case dehd. Pet Milk
Fills respiration requirement	1	Two 100 lb. tanks oxygen
One of principal means of finding directions	3	Stellar map
CO <sub>2</sub> bottles for self-propulsion across chasms, etc.	9	Life raft
Probably no magnetized poles; thus, useless	13	Magnetic compass
Replenishes loss by sweating, etc.	2	Five gallons of water
Distress call when line of sight possible	10	Signal flares
Oral pills or injection medicine valuable	7	First aid kit w/injection needles
Distress signal transmitter, possible communication with mother ship	5	Solar-powered radio

FIG. 1. NASA Decision-Making Problem, leader key.

of favorability for the leader, with the appropriate leader style dependent on the degree of favorability (Fiedler, 1967).

Favorability is a direct function of three contingency variables which are in decreasing order of importance: (a) leader-member relations (good or poor); (b) task structure (structured or unstructured); (c) leader-position power (strong or weak). For example, based on empirical evidence, Fiedler would predict that the authoritarian leader would be most effective in a favorable environment (good leader-member relations, structured task, high-position power) or in an unfavorable environment (poor leader-member relations, unstructured task, weak-position power). Concomitantly, democratic leadership style would be appropriate and most effective for moderately favorable environments (good leader-member relations, unstructured task, weak-position power).

Effective leadership is desired and sought after by all formal organizations particularly by those faced with chronic crisis-oriented situations such as the police and the military. Such concepts as the life Cycle Theory and Fiedler's Contingency Model offer potential contributions for improvement of leadership techniques through identification of appropriate behaviors leading to successful leadership in specific environments and situations.

This experiment attempted to contrast effects of authoritarian and democratic leader-

ship styles through manipulation of Fiedler's contingency variables under a limited time constraint. Specifically, the purpose of this experiment was to determine whether democratic or authoritarian leadership was more effective under the conditions of high- or low-leader task accuracy.

#### METHOD

##### Subjects

This experiment was conducted in two parts. Each study employed different samples of subjects. The initial study utilized 48 four- or five-man groups of United States Air Force Academy cadets, while the replication conducted 1 year later utilized 32 groups of four or five United States Air Force Academy cadets. All subjects were male sophomores or juniors between the ages of 19 and 23, and enrolled in an advanced leadership course. These subjects were considered appropriate for this type of experiment because of their willingness to cooperate and to accept perceived legitimate authority in an academic environment.

##### Materials

Each subject received an individual copy of the National Aeronautic Space Administration (NASA) Decision-Making Problem. The problem consisted of instructions as given in the procedure section below and a listing of the survival items appearing in Figure 1. Group leaders were issued a copy of the group summary form. In addition to the list of survival items depicted in Figure 1, this form contained spaces for recording the predictions of each group member and the final group predictions. At least 24 hours prior to each trial, a copy of Figure 1 containing the solution was also given to each leader for his private use.

### Procedure

In order to provide a realistic, structured task with a known solution, all subjects were initially administered individually the NASA Decision-Making Problem. The following instructions were read and distributed to the assembled subjects prior to the start of each trial:

You are a member of a space crew originally scheduled to rendezvous with a mother ship on the lighted surface of the moon. Due to mechanical difficulties, however, your ship was forced to land at a spot some 200 miles from the rendezvous point. During reentry and landing, much of the equipment aboard was damaged; and since survival depends on reaching the mother ship, the most critical items available must be chosen for the 200-mile trip. Below are listed the 14 items left intact and undamaged after landing. Your task is to rank order them in terms of their importance in allowing your crew to reach the rendezvous point. Place the number 1 by the most important item, the number 2 by the second most important item, and so on, through number 14, the least important. Work alone. Do not compare answers. You have 10 minutes to complete the task. All solutions will be compared for accuracy upon completion.

Upon expiration of the 10-minute limit for individual estimates, subjects were randomly assigned to groups of four or five. Random assignment was accomplished by arranging the cadets assigned to a class alphabetically and systematically assigning every fourth or fifth subject to a particular group. Subsequently, leaders were publicly appointed, given group summary forms and told to assume responsibility for guiding the group to a consensual agreement on the rank order of importance of the survival items. A 30-minute time limit was imposed for completion of group activities.

Unknown to the other subjects, leaders had been briefed prior to the trials, given the correct solution to memorize as shown in Figure 1, and instructed to adopt certain behavioral roles during the consensual process. The specification of these roles was crucial to the rationale of this experiment, for adherence to a specified behavioral role insured that the type of behavior desired from designated leaders would be elicited. Half the leaders were instructed to use an authoritarian leader style. Of these, one half were told to sway the group to the most accurate solution and the other half were instructed to sway the group to the least accurate solution possible. The other half of the leaders were briefed to utilize the democratic leader style. One half of these were also told to sway their groups to the most accurate solution, and one half were to attempt to achieve the least accurate solution.

For the purposes of this study, the leadership styles were defined as follows: The authoritarian leader assumes and exercises complete control of the group in determining task structure, methodology and decision making toward completion of the task. Authoritarian leader behavior emphasizes task com-

pletion above all other considerations. Conversely, the democratic leader shares responsibility for determining task structure, methodology, decision making, and task completion with the other members of the group. Consequently, democratic leader behavior is directed toward maintenance of harmonious interpersonal relationships as the primary means for task achievement.

In keeping with the above definitions, authoritarian leaders were instructed to maintain control of the group, argue absolutely for acceptance of their solutions, ignore any alternative solutions incompatible with their own, and to attempt to make all final decisions with complete autonomy. Democratic leaders were instructed to serve as facilitators with the primary aims of minimizing group conflict and enabling every group member's ideas to be aired and considered. Additionally, democratic leaders were advised to aid the group in achieving a consensual decision as opposed to exercising autonomous authority.

Prior to this study, the subjects had received considerable instruction and experience in leadership techniques both in the classroom and through their military training. Because of these factors the above definitions and behaviors were easily recognized and understood.

Thus, democratic and authoritarian leadership styles were combined with levels of high and low leader accuracy in order to measure differences in group problem solving accuracy. Average group absolute error scores were used for treatment comparisons, while grouped average error scores of individual estimates were used to insure that the four treatment conditions were equivalent in terms of initial subject accuracy.

### RESULTS

Table 1 depicts the summary of results for both studies. The pattern of results were similar for both the initial study and the replication. Error scores for authoritarian-led groups produced both the highest and lowest accuracy. This was directly related to the degree of accuracy employed by the leaders. Democratic-led groups produced intermediate accuracy levels. Democratic groups with high-accuracy leaders were only slightly different numerically from democratic groups led by leaders with low accuracy. A *t* test was used to analyze differences between means of the same leader styles for all four treatment conditions. No significant differences were found between the means for the two studies ( $p < .05$ ).

In order to determine if the mean differences between different leader styles were significant, a one-way analysis of variance was

TABLE 1  
SUMMARY OF MEANS AND STANDARD DEVIATIONS

Leader style	Initial Study (1970)			Replication (1971)			Diff.	df	<i>t</i>
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>			
TYPE I Authoritarian—high accuracy	12	10.42	9.04	8	7.25	8.21	3.17	18	0.76
TYPE II Democratic—high accuracy	12	20.67	6.40	8	21.38	5.53	.71	18	0.24
TYPE III Authoritarian—low accuracy	12	33.33	7.15	8	38.50	6.77	5.17	18	1.64
TYPE IV Democratic—low accuracy	12	26.08	7.04	8	27.00	7.03	.92	18	0.27

conducted for each study. Significant differences between leader styles were found in both studies ( $p < .001$ ,  $F = 18.6$  and  $F = 14.23$ ). Additionally, the Neuman-Kuels, a posteriori method for testing differences between means, was conducted. The performance of Leader Style I was found to be significantly better than the other three styles ( $p < .01$ ). Additionally, Leader Styles I, II and IV were all found to be significantly more accurate than Leader Style III. It should be noted that Leader Styles II and IV, the high and low accuracy democratic-led groups, did not differ significantly from one another ( $p > .05$ ). The results were the same for both studies.

The data indicate that authoritarian-led groups produced the highest or lowest accuracy as a direct consequence of the degree of accuracy employed by the leaders while democratic-led groups produced intermediate accuracy levels which were statistically equivalent despite the extreme differences in the accuracy levels of the leaders.

A summary of results for the average individual error scores is depicted in Table 2. Individual error scores were averaged for each group. Group means were then averaged for each of the four leader styles. For both studies, these means appear to differ only slightly numerically across the four treatment conditions.

TABLE 2  
MEANS OF AVERAGE INDIVIDUAL ERROR SCORES

Leader style	Initial Study (1970)			Replication (1971)		
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>
TYPE I Authoritarian—high accuracy	12	32.16	13.39	8	33.80	18.60
TYPE II Democratic—high accuracy	12	34.01	17.09	8	36.30	17.21
TYPE III Authoritarian—low accuracy	12	35.90	15.94	8	32.70	19.69
TYPE IV Democratic—low accuracy	12	31.84	13.64	8	35.90	17.09

In order to determine if these differences were significant, *F* tests were conducted. In both cases the overall *F* was found not significant at the .10 level (*F* = 2.19). It was concluded that the average individual error scores were equivalent for the four treatment conditions. Thus, the differences in consensual decision scores found between treatment conditions in both the initial and replication studies could not be attributed to systematic differences in average individual error scores.

Each group was observed by an experimenter in order to evaluate group processes during the exercises. The observers recorded their subjective descriptions of interpersonal activities indicating degrees of conflict, cohesiveness and communication flow. Following completion of the consensual process, discussion periods were held to collect additional verbal data from leaders and group members. All results were content analyzed by the five experimenters to determine whether there were gross behavioral differences between the four treatment conditions. Typically, authoritarian-led groups were characterized by aggressive and hostile verbal acts between leader and group while democratic-led groups were characterized by lack of hostility and aggression with cooperation and harmony in much evidence. Shouting and disagreements increased in intensity while flow of communication decreased between group members and leaders in authoritarian-led groups as the time limit was approached. This increase in verbal activity and concomitant decrease in communication flow was not observed in the democratic-led groups.

#### DISCUSSION

The results of this experiment support some of the empirical data surveyed earlier (Fiedler, 1967; Hersey & Blanchard, 1969; Olmstead, 1967). The activities of authoritarian-led groups were characterized by conflict and hostility, especially the high-accuracy groups which suffered from marked verbal clashes, aggression toward leaders, and a high number of disagreements despite the high accuracy of the leaders. The experimenters also noted that the total number of communication acts between leader and group members typically diminished during the con-

sensual process. These phenomena support the findings of Lewin, Lippit, and White (1939). However, some of the hostility and aggression observed here could be attributed to the effects of normative behavior directed by subjects against peers who exercise authoritarian leadership in an academic environment.

The data support the predictions of the Contingency Model in that authoritarian leadership was most productive under conditions of good leader-member relations, a structured task and strong leader position power. In terms of goal achievement, which is synonymous with group accuracy for our purposes, the data indicate that highly accurate authoritarian leaders were most successful, authoritarian leaders with low accuracy were least successful, and democratic leaders produced moderate degrees of goal accomplishment which appear to be independent of leader accuracy.

Fiedler's contention that personality tendencies limit one's opportunities for successful leadership tended to be supported by observations of the experimenters and analysis of the verbal comments of the leaders. The leaders in many cases did play roles that were in dissonance with their personalities because of the random assignment of leaders to the four treatment conditions. Many were uncomfortable in their role playing. This was particularly true of the highly accurate authoritarians who perceived much group dissatisfaction with their leadership despite highly effective goal accomplishment. This could have implications for long-range effectiveness. For example, several leaders felt that prolonged role playing over a period of time might be deleterious to their psychological state of adjustment because of conflict between personal goals and the environmental requirements.

With regard to the variables of the Contingency Model, the public appointment of leaders assured them strong position power as evidenced by their ready acceptance as leaders by the group members. The specific nature of the task, the concreteness and familiarity with the problem stimuli and the explicitness of the instructions all aided in the structuring of the task. Additionally, initial good

leader-member relations were indicated by lack of hostility, intragroup harmony and willingness of the subjects to accept the problem and the leader's authority.

The deterioration of leader-member relations during the authoritarian consensual processes could indicate a possible weakness in the Contingency Model. One might question whether the contingency variables manipulated here are as static as the Model seems to imply. For example, during organizational activities, is it not possible for unstructured tasks to become structured, or for leader position power to increase or decrease? Also, the Model makes no provision for consideration of the group's maturity level or for the temporal variable which could prove consequential for short term effects on productivity and group harmony. Finally, the Contingency Model does not provide for any combination of authoritarian and democratic leader styles as does the Life Cycle Theory.

The successful accomplishments of the leaders despite random selection for role playing tends to support the notion of adaptability as espoused by the Life Cycle Theory. The ability to modify behavior as a consequence of environmental and situational requirements was demonstrated here. However, the long-term effects of personal conflict developed through dissonance of personality and organizational requirements remain in doubt. Perhaps combining Contingency Model concepts for long range effects with Life Cycle Theory applications for short-term effects could be fruitful. Experimental study would seem desirable.

A possible weakness of the study was that both the collection and analysis of the subjective data were performed by the same individuals. This had the advantage of interpretation based upon observation but

could potentially have resulted in some contamination related to personal bias.

Finally, the data tend to support Holloman and Hendrick (1970) who found that group consensual decisions were more accurate than the average of individuals on the same problem-solving task used in this study. It is interesting to note (see Tables 1 and 2) that authoritarian leaders with low accuracy provide the sole exception to their findings. Comparisons of the within-treatment means in both of the present studies with the corresponding average individual error scores indicate that these leaders may significantly distort the group average error score to a level comparable with the average of individual errors, thus negating the value of consensus in terms of productivity.

#### REFERENCES

- FIDLER, F. A theory of leadership effectiveness. New York: McGraw-Hill, 1967.
- HERSEY, P., & BLANCHARD, K. H. Management of organizational behavior. Englewood Cliffs, N. J.: Prentice-Hall, 1969.
- HOLLANDER, E. P., & JULIAN, J. W. Contemporary trends in the analysis of leadership process. *Psychological Bulletin*, 1969, 71, 387-397.
- HOLLOWAY, C. R., & HENDRICK, H. W. Individual versus group effectiveness in solving factual and nonfactual problems. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 1970.
- KORMAN, A. K. Industrial and organizational psychology. Englewood Cliffs, N. J.: Prentice-Hall, 1971.
- LAMBERT, W. W., & LAMBERT, W. E. Social psychology. Englewood Cliffs, N. J.: Prentice-Hall, 1964.
- LEWIN, K., LIPPIT, R., & WHITE, R. K. Leader behavior and member reactions in three 'social climates.' *Journal of Social Psychology*, 1939, 10, 271-279.
- OLMSTEAD, J. A. The skills of leadership. *Military Review*, March 1967, 62-70.
- YUKL, G. A., & WEXLEY, K. N. Readings in organisational and industrial psychology. New York: Oxford University Press, 1971.

(Received August 30, 1971)

## EFFECTS OF EMOTIONAL AROUSAL ON THE USE OF SUPERVISED COERCION WITH BLACK AND UNION EMPLOYEES

DAVID KIPNIS,<sup>1</sup> ARNOLD SILVERMAN, AND CHARLES COPELAND

*Temple University*

It was predicted that situations capable of inducing negative affective states among supervisors would promote the use of coercion by supervisors. First-line supervisors described an incident in which they used delegated powers to correct subordinate behavior. Analysis of these incidents revealed that supervisors used more coercion with black than white subordinates, and with union than nonunion subordinates. It is assumed that in both instances, heightened emotional responses caused by prejudice in the case of black subordinates and resistance to orders in the case of union members induced the use of coercion.

When individuals disagree, the use of coercive power by either party makes the re-establishment of harmonious relations very difficult. Studies of the use of threat and coercion in bargaining and conflict situations suggests that the control of coercive power tempts the individual to impose his own wishes on others (Deutsch & Krauss, 1960). Contrarily, less powerful individuals tend to resist compromising with more powerful sources, if such resistance is not too costly (French & Raven, 1959; Swingle, 1970). These temptations to use coercive power as a means of imposing one's will, as well as the counterinclinations to resist, if possible, intensify conflict and interpersonal hostility.

The present study is concerned with conditions that influence the use of coercive power. It is one in a series concerned with the more general question of how power is used within organizational settings. The focus has been on the first-line supervisor, who not only may use personal bases of power (e.g., persuasive power, physical strength, personal charm, etc.) for influencing subordinates but also has limited access to a range of institutional powers, as these are associated with his office. These latter powers extend the supervisor's potential for influencing others. Our studies have attempted to identify conditions that influence the supervisor's choice of powers when attempting to change the behavior of subordinates.

<sup>1</sup> Requests for reprints should be sent to David Kipnis, Department of Psychology, Temple University, Philadelphia, Pennsylvania 19122.

Berkowitz (1970) has reported a series of studies that point to the importance of emotional arousal as a prerequisite for aggressive behavior. A review of the data from our previous studies also suggest that emotional arousal among supervisors was present when these leaders relied on threats and coercion. In field studies (Kipnis & Cosentino, 1969; Kipnis & Lane, 1962) as well as in laboratory simulations of business (Goodstadt & Kipnis, 1970; Kipnis & Vanderveer, 1971), threats and coercion were used when subordinates manifested hostility and poor work attitudes, or when supervisors were uncertain about what to do, or were overburdened by the requirement that they supervise large numbers of men. What these situations appear to have in common is that they induce negative emotional states within the appointed leaders. If this belief is correct, it could be expected that in other situations that have the potential for evoking negative emotional states, there should be reliance upon coercion.

The present study is concerned with two such situations. The first is concerned with the extent to which coercion is used by supervisors when dealing with white and black subordinates. If supervisors feel emotional antipathy toward black subordinates, then coercion should be associated with attempts to change these subordinates' behavior. The second situation likely to evoke negative affect is the presence of an active union. Our prior field study was conducted in a non-union setting, where the ability of subordinates to resist supervisors' influence was relatively weak. The presence of a union, how-

ever, approaches a situation of bilateral power (Deutsch & Krauss, 1960). With the support of the union, subordinates can actively resist supervisors' influence. In turn this could provoke resentment or anger, or even feelings of frustration and helplessness, among supervisors. Under these conditions of emotional arousal, given that supervisors have access to coercive powers, we predict that they will use them.

More specifically, then, the purpose of the present study is to investigate the use of coercive power by supervisors among black and white employees, and between union and nonunion employees. From what has been said above, it is expected that more coercion will be used with black and union employees.

### METHOD

The study was conducted in an eastern steel plant in which all hourly paid employees were union members. The union strongly represented its members at bargaining and grievance sessions. The subjects of this study were 66 first line supervisors, 57 who were white and 9 who were black.

In previous field studies the use of power was measured through a critical incident technique. In this technique, supervisors were asked to describe an incident in which they corrected the substandard behavior of subordinates.

A content analysis of these incidents provided data on both the nature of the subordinate's problem and the steps the supervisor took to correct the subordinate's behavior. This latter information provided data concerning the range of powers available in that given situation. Because content analysis involved difficulties in coding and interpretation, a more objective procedure was used in the present study.

From preliminary interviews with supervisors and their superiors, a checklist was constructed of power-based actions available to supervisors when attempting to correct subordinates' behavior. This checklist contained 27 items that could be classified into four categories of power usage, and one category reflecting the supervisor's attempts to get help from others in dealing with the subordinate. An additional item was provided for those supervisors who did nothing about the problem subordinate. The items, grouped by area, are given in Table 1. Instructions were to check as many items as applied. The score for a given area was the total number of items checked.

Supervisors were either contacted by mail or in a supervisory training session. In either instance, the questionnaire was described as a university project, supported by the company, concerned with studying the range of problems encountered by supervisors. The questionnaire was anonymous, although it was coded to identify the race of the respondent.

TABLE 1

CHECKLIST ITEMS FOR EVALUATING POWER USAGE BY SUPERVISORS

Item
Persuasive power
1. I asked him what the problem was. 2. I explained to him how his behavior was causing trouble.
Expert power
3. I took some time to show him what he was doing wrong. 4. I kept close watch on him to make sure he was doing his job. 5. I tried to set an example for him by my own actions. 6. I told him he should use one of his fellow workers as an example.
Ecological change
7. I gave him work he was better at. 8. He was transferred.
Coercive power
9. I threats and reprimand 10. I chewed him out 11. I gave him a verbal warning 12. I threatened to give him a written warning 13. I ignored him while being friendly to everyone else. 14. I kept riding him.
Reduction in work privileges
15. I scheduled him to work hours he didn't like. 16. I gave him work he didn't like. 17. I put him in a work area he didn't like. 18. I gave him a written warning. 19. I took steps to suspend him. 20. I recommended that he be brought before the Disciplinary Committee. 21. He was suspended. 22. He was fired.
Seeking advice from others
23. I talked it over with my supervisor. 24. I talked it over with the other foremen. 25. I talked it over with some of the problem-employee's co-workers. 26. I sought help from another department.
Avoidance of action
27. There was nothing I could do.

The first part of the questionnaire asked the supervisor to describe an incident that occurred within the past year in which he had to correct the below-average performance of one of his employees. This information was content analyzed into four categories describing the type of problem encountered. The first category involved *problems of work*—for instance, the supervisor may have written, "This employee was very slow to catch on when you gave him orders and installed equipment improperly." The second category involved *problems of poor attitude*—for example, "An employee felt that because he was the oldest he did not have to help on a team job."

The third category involved the breaking of company rules, coded as *discipline*—such things as lateness, drinking, and stealing were included here. Finally, a fourth category included all coded incidents that could be classified as involving at least two of the previously mentioned categories, for example, a subordinate who did a poor job of repairing an oxygen connector because of a dispute as to whose job it was. This category was coded as *complex problems*, consisting of combinations of work, attitude, and discipline. Working independently, two coders agreed in their classification of the type of problem in 92% of the incidents. Where disagreement occurred, they were discussed until agreement was reached.

The second part of the questionnaire asked the supervisors what they did about the problem reported and was followed by the mentioned checklist in Table 1. Information was also asked on the race of the subordinate and his length of company employment.

## RESULTS

We will report in detail only the data concerned with the use of coercive power (i.e.,

items 9–22). However, an analysis of factors influencing the use of the complete range of powers essentially replicated findings from the earlier field study (Kipnis & Cosentino, 1969). That is, the nature of the problem, span of control, and complexity of the problem significantly influenced the supervisor's choice of means of influence. Thus, problems of work evoked expert powers, problems of attitude or discipline evoked coercion, supervisors directing large numbers of men relied on coercion, and complex problems evoked the use of a larger number of powers than simple problems.

In the present sample, years of experience of the supervisor was not related to his use of power, as was previously found. Advice of others (items 23–26) was sought significantly more frequently when the employee's problem was coded as complex rather than simple.

Despite the large number of alternatives pertaining to the use of coercion, this form of power was not the most popular. Ninety-seven percent of all supervisors checked at least one item pertaining to persuasion, 83% checked at least one item pertaining to the use of expert power, 56% checked at least one item pertaining to the use of coercion, and 17% checked at least one item concerned with ecological change. Seventy-one percent of the supervisors sought the advice of others.

Table 2 gives the distribution of the kinds of problems encountered by supervisors. The problems reported were divided relatively equally between work, attitudes, discipline, and more complex problems. For comparison purposes, the distribution of problems reported by nonunion supervisors previously studied (Kipnis & Cosentino, 1968) is also shown in Table 2.

Differences between these two distributions of problems were evaluated by a chi-square test. It can be seen in Table 1 that supervisors of union employees reported that their subordinates manifested significantly more attitudinal problems ( $p < .01$ ), more complex problems ( $p < .10$ ), and fewer problems of poor work ( $p < .01$ ) than did the previously studied nonunionized sample. However when the complex problems were broken into those that also included work

TABLE 2  
DIFFERENT CLASSES OF SUBORDINATE PROBLEMS  
REPORTED BY SUPERVISORS

Problem	Union sample in percent	Non- union sample <sup>a</sup> in percent	<i>p</i>
Work	28	47	<.01
Attitude	26	8	<.01
Discipline	20	27	ns
Complex problems	26	17	<.10
	100	100	
Number of supervisors	66	131	
Total percentage of supervisors mentioning:			
Work problems	54	62	ns
Attitude problems	47	18	<.01
Discipline	30	36	ns

<sup>a</sup> Data abstracted from Kipnis, D., and Cosentino, J. *Journal of Applied Psychology*, 1969, 53, 460–466.

and attitude, as shown in Table 2, only attitude problems distinguished between the union and nonunion sample. As was already mentioned, these findings may be interpreted to mean that the presence of a union leads to less compliance among subordinates, more conflict between subordinates and supervisors, and hence more reports of poor attitude problems by supervisors.

Our prior research revealed that when supervisors encountered problems of poor attitude or discipline, they invoked coercive power. Since more attitudinal problems were reported in the present study, it could be expected that more reliance would be placed on coercion than in the previous study. Unfortunately, because of differences in methodology between the present study and the previous one (i.e., content analysis vs. checklist), it was not possible to statistically compare the frequency of use of coercion in the two samples. However, inspection of the data strongly suggests that supervisors in the present study used coercion more frequently than did supervisors of nonunion men. Table 3 shows the percentages of supervisors of union and nonunion men who used each of the various forms of coercion. To make the data consistent with the prior study, the item "man fired" is presented separately. It can be seen that more supervisors in the present sample reported using threats and reprimands (38% vs. 16%), reductions in work privileges (6% vs. 1%), and administrative punishments (official warnings, reports, and suspensions) (19% vs. 7%), than did supervisors of nonunion men. On the other hand, there was a slight trend for nonunion subordinates to be fired more often (3% vs. 8%), suggesting, perhaps, that the presence of a union places restraints on the use of this form of coercive power. While these differences in the use of coercion were in the predicted direction, as was mentioned above, the differences should be treated with caution because of methodological differences in the collection of the data.

The next problem examined was whether the race of the subordinate influenced the use of coercive power. For this analysis only three white supervisors were used. Three supervisors did not state the race of their subordinates and were dropped from the

TABLE 3  
DIFFERENT TYPES OF COERCIVE POWER  
USED BY SUPERVISORS<sup>a</sup>

Coercive Device	Union sample in percent	Non-union sample <sup>b</sup> in percent
Threats and reprimands	38	16
Reduced privileges	6	1
Administrative punishments (Written warnings)	19	7
Man fired	3	8
Number of supervisors	66	131

<sup>a</sup> The percentages within each category are based on the number of supervisors who checked at least one of the items comprising that category.

<sup>b</sup> Data abstracted from Kipnis, D., and Cosentino, J. *Journal of Applied Psychology*, 1969, 53, 460-466.

analysis. Among the remaining white supervisors, 19 reported an incident involving a black subordinate and 35 reported an incident involving a white subordinate. There was no significant difference in the length of time the black or white subordinates had been employed on the job, although more of the black subordinates (53%) than white subordinates (42%) had been employed less than a year ( $\chi^2 = .68$ ,  $p$  is ns). Further, there were no differences reported in the kinds of problems manifested by white and black subordinates. The distribution of problems was practically identical.

Despite these similarities, it was found that supervisors invoked their administrative coercive powers more frequently with black than white subordinates. The average frequency of use of administrative punishment among white subordinates (i.e., the sum of administrative coercion alternates checked on the checklist divided by the number of supervisors) was .17. The corresponding figure among black was .63. This difference was significant beyond the .05 level ( $F = 4.05$ ,  $1/52$  df). Stated another way, 32% of the black subordinates and 14% of the white subordinates were fired, suspended, given written warnings, or recommended for disciplinary actions.

There were no differences between white and black subordinates in the use of threats and reprimands or in the use of reductions

in work privileges, although slightly more supervisors invoked this latter category of coercion with black than with white subordinates (11% vs. 5%). While perhaps obvious, these findings suggest that discriminatory treatment of blacks does not exist solely at the level of selection, but may be detected in supervisory behavior as well. Supervisors appear to punish infractions of black employees more harshly by invoking administrative punishment than when the same infractions were made by white employees.<sup>2</sup>

### DISCUSSION

The use of coercive power is pervasive in our society. Its expression takes many forms. In the present study the threats used by supervisors were primarily economic in nature and had to do with loss of jobs, wages, chances for advancement, and the like. The fact that these threats were used more frequently among black employees and union members is consistent with the hypothesis that situations capable of inducing negative affective states would promote the use of coercion. In the case of union members, the emotional feelings of supervisors were presumably aroused by employees showing resistance to supervisors' orders. In the case of black subordinates, it is believed that hostility toward these subordinates caused an increased reliance on coercion. Obviously a more direct test of these beliefs would involve obtaining measures of racial attitudes of supervisors in the latter instance, and measures of subjective stress in the former, and relating these measures to the use of coercion.

Berle (1967) has pointed out that in organizational settings, individuals with access to power frequently find that the obligations of power force them to behave in ways that conflict with their personal values. Because of their involvement in the organization, individuals find they must use any and all institutional powers that are available to pro-

tect and extend corporate functioning, despite any feelings of personal misgivings. Within this context, we view supervisors' response to union employees as representing a form of role-induced use of power. That is, given involvement in organizational goals, supervisors may have felt annoyed or angered over their subordinates' poor attitudes, because these attitudes blocked the attainment of organizational objectives or represented what supervisors considered to be failure by employees to accept their legitimate role obligations. Hence supervisors felt obliged to invoke coercion to protect the organization. In this instance then, the use of power represented the individual fulfilling his perceived role obligations.

Berle (1967) has also pointed out that access to institutional power can be used by individuals in the service of personal goals, as contrasted with institutional goals. According to Rogow and Lasswell (1963), one manifestation of the corrupting influence of power is that it tempts individuals to use institutional power to satisfy personal rather than institutional needs. The excessive coercive power used among black subordinates appears to be an instance in which institutional powers were used for personal reasons. While perhaps not only aware of the dynamic reasons involved, bias apparently caused supervisors to overreact to the substandard behavior of their black subordinates. As such, the punishments invoked served to gratify personal needs of supervisors, rather than organizational needs.

In all instances, access to institutional powers allows the individual to extend his influence over others. While powerless individuals may "turn the other cheek," this conciliatory gesture is far less likely to happen when institutional powers of a coercive kind are possessed. In some instances the individual may feel forced by his loyalty to the institution to retaliate, regardless of the harm done to the target of power. In other instances, as was suggested above, the distinction between institutional goals and personal needs becomes blurred, so that the individual diverts institutional powers to satisfy his own wants.

<sup>2</sup> Readers may note the similarity of the present findings to charges made in recent Congressional hearings investigating racial riots aboard U. S. Navy ships. Here also black sailors testified that they were punished more harshly than white sailors for the same offenses.

While the present study investigated these two uses of power among first-line supervisors, case studies clearly reveal that similar uses of power can be found at higher levels of management as well. For example, Galbraith (1967) and Fuller (1962) both report instances in which managers of large corporations felt obliged to invoke institutional powers of essentially a coercive kind against weaker targets, despite the fact that this use of power violated laws and the general welfare of the public. Contrarily, Jay (1967) describes instances in which institutional powers were used by managers to satisfy personal ambitions. These studies suggest the importance of studying the interplay between the individual and the institutional powers he controls at all levels, as these powers are used to influence behavior within the institution and events outside the institution.

## REFERENCES

- BERKOWITZ, L. The contagion of violence. In W. S. Arnold & M. M. Page (Eds.), *Nebraska symposium on motivation*. Lincoln: University of Nebraska Press, 1970.
- BERLE, A. A. *Power*. New York: Harcourt, Brace and World, 1967.
- DEUTSCH, M., & KRAUSS, R. M. The effects of threat upon interpersonal bargaining. *Journal of Abnormal and Social Psychology*, 1960, 61, 181-189.
- FRENCH, J. R. P., JR., & RAVEN, B. H. The bases of power. In D. Cartwright (Ed.), *Studies in social power*. Ann Arbor: University of Michigan, 1959.
- FULLER, J. G. *The gentlemen conspirators*. New York: Grove Press, 1962.
- GALBRAITH, K. J. *The new industrial state*. Boston: Houghton Mifflin, 1967.
- GOODSTADT, B., & KIPNIS, D. Situational influences on the use of power. *Journal of Applied Psychology*, 1970, 54, 201-207.
- JAY, A. *Management and Machiavelli*. New York: Holt, Rinehart & Winston, 1967.
- KIPNIS, D., & LANE, W. Self-confidence and leadership. *Journal of Applied Psychology*, 1962, 46, 291-295.
- KIPNIS, D., & COSENTINO, J. Use of leadership powers in industry. *Journal of Applied Psychology*, 1969, 53, 460-466.
- KIPNIS, D., & VANDERVEER, R. Ingratiation and the use of power. *Journal of Personality and Social Psychology*, 1971, 17, 280-286.
- ROGOW, A. A., & LASSWELL, H. S. *Power, corruption and rectitude*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.
- SWINGLE, P. G. *The structure of conflict*. New York: Academic Press, 1970.

(Received July 23, 1971)

## THE INFLUENCE OF SEX-ROLE STEREOTYPES ON EVALUATIONS OF MALE AND FEMALE SUPERVISORY BEHAVIOR

BENSON ROSEN<sup>1</sup> AND THOMAS H. JERDEE

*Graduate School of Business Administration, University of North Carolina*

This investigation examined the way sex-role stereotypes—perceptions and expectations of what is appropriate behavior for males and females— influence evaluations of male and female supervisory behavior. Undergraduate students and bank supervisors were asked to read one of six versions of a supervisory problem (with either a male or female supervisor and with either male, female, or mixed subordinates) and to evaluate the effectiveness of four supervisory styles. Results indicated that sex-role stereotypes do influence evaluations of supervisory effectiveness for some, but not all of the supervisory styles. Findings are discussed in terms of the potential negative consequences of sex-role stereotypes for supervisory behavior.

Traditionally, women have been limited mainly to clerical, operative, nursing, teaching, and social service occupation (Kreps, 1971) and most research on organizations has been based on the assumption that managerial positions are the special province of males. There have been few scientific studies of women managers. In the future, however, it seems reasonable to assume that women might be employed in almost any managerial position currently staffed by men. Factors contributing to this expected change in aspirations of women are: (a) changing cultural values concerning the role of women in society, (b) federal legislation banning sex discrimination in employment practices (specifically, Title VII, Civil Rights Act, 1964), (c) increasing opportunities for women to acquire advanced education and training, and (d) the increasing number of young women with work experience and no small children.

The introduction of women into managerial ranks represents a new challenge for educators, employers, and organizational psychologists, who must become increasingly concerned with the special characteristics of women that might be of relevance to their performance in supervisory roles and with new questions involving male-female interactions. A matter of particular concern is the pos-

sible clash between prevailing expectations regarding the appropriate behavior for women as females and expectations regarding the supervisory role.

Several writers have depicted differential societal expectations for male and female behavior. Tyler (1965) for example, has suggested that women are expected to be sympathetic, humanitarian, compassionate, and dependent on others. Expectations for females also include nonaggression (Hilgard & Atkinson, 1967), spiritual values, artistic inclinations and concern for the welfare of others (Miner, 1965). Conversely, a behavioral orientation toward power, initiative, and prestige is frequently viewed as more appropriate for males (Miner, 1965). The present investigation is concerned with how these general societal expectations regarding male and female behavior influence more specific occupational role expectations for male and female supervisory personnel in formal organizations.

There is some indirect evidence that supervisory role expectations are applied with equal force in judging male and female supervisors and that women are judged as less likely to meet these expectations, presumably because of the clash with generally accepted sex-role expectations. Studies by Klein (1950) and Scheinfeld (1944) document a tendency toward prejudicial evaluation of women's work by men. Gilmer (1961) found that over 65% of male managers believed that women would be inferior to men in supervisory jobs.

<sup>1</sup> Requests for reprints should be sent to Benson Rosen, Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514.

They believed that women have higher absenteeism than men, are more neurotic than men, and have more work-related problems than men. More recently, it has been shown that the way women behave on the job rather than the way they perform the technical operations of their positions is the chief determinant of their acceptance as administrators (Gilmer, 1971).

The tendency to devalue women's performance is not limited to men. Goldberg (1968) has shown that women are also quite prejudiced in their evaluation of the intellectual and professional competence of other women. Goldberg asked college students to evaluate journal articles that were attributed in some cases to a male author and in other cases to a female author. Evaluations of articles that were attributed to female authors were lower than evaluations of the same articles attributed to male authors. In a second study concerned with the evaluation of performance by women (Pheterson, Kiesler, & Goldberg, 1971), it was concluded that women who are striving for accomplishment are judged less favorably than men, but women who have successfully accomplished work are evaluated as favorably as men.

The present study used an approach similar to Goldberg's in order to investigate the effects of a supervisor's sex on people's evaluations of his or her potential effectiveness. As in Goldberg's study, some subjects were presented with a situation involving a female supervisor and some with a situation involving a male supervisor. Subjects were not aware that a comparison of male and female supervisors was involved. Rather, the task as it appeared to them was simply to evaluate the propriety and potential effectiveness of four alternative supervisory approaches that were being considered by the supervisor depicted in the case description.

Our basic research hypotheses were concerned with the effects of the supervisor's sex on our research subjects' evaluations of the supervisor's potential effectiveness. First, we hypothesized that evaluations would generally be higher for male supervisors because culturally expected "female" behavior would be viewed as conflicting with role demands for supervisors.

A second hypothesis was that there would also be a sex-style interaction effect, with female supervisors judged as more likely to succeed with certain supervisory approaches and male supervisors with others. This effect would depend on the degree of congruence between the particular supervisory approach involved and the judge's sex-role stereotype—his perception of what is generally considered appropriate behavior for each sex.

We also hypothesized that the aforementioned effects would occur regardless of the sex of the person making the evaluation and regardless of his or her current employment status (college student or bank supervisor).

## METHOD

### Subjects

Subjects were drawn from two populations. During the spring semester of 1971, 134 male and 24 female undergraduate business students participated in the study. A few weeks later, 83 male and 15 female banking supervisors attending a management institute at the University of North Carolina served as subjects. Thus, a total of 256 subjects were asked to make evaluations of male or female supervisory behavior.

### Experimental Design

The two manipulated variables were sex of the supervisor (male or female) and sex of the subordinates (males, females, or both males and females). In addition, the sex of the subject and the subject's status (student or bank supervisor) were recorded. Each subject participated in only one experimental condition, and assignment of subjects to conditions was completely randomized.

### Procedure

Subjects were presented with experimental materials in their regular classrooms as part of a class exercise. Each subject was issued one of six versions of a booklet entitled *Supervisory Styles*, and instructed to read the following directions:

We would like to get your opinion about the appropriateness and effectiveness of various supervisory styles. Please read the following supervisory problem and indicate your opinions on the scales provided.

Ruth (Ralph) Brown is 41, married and lives in a downtown apartment. She (He) has had considerable experience in office management work.

Mrs. (Mr.) Brown was recently hired as office manager for the Ordinal Oil Company, a rather

small mid-western distributor. In this job she (he) is in charge of twelve male (twelve female) (six male and six female) clerical employees.

The case portrays Mrs. (Mr.) Brown as troubled by the high absenteeism and poor work performance of the clerical staff. Four possible courses of action are considered by the supervisor as means of maintaining work standards among the clerical staff:

1. Firmly advise her (his) subordinates that they would be discharged unless there is significant improvement in their work (*threat*).
2. Advise her (his) subordinates that forthcoming recommendations for salary increases would depend on improved performance (*reward*).
3. Approach her (his) subordinates in a friendly way and asks them to help her (him) by improving their performance (*friendly-dependent*).
4. Offer to help subordinates with any problem that might impede their performance (*helping*).

*Subjects* were then instructed to evaluate each of the alternatives on the following three bipolar semantic differential scales: bad-good, improper-proper, and ineffective-effective. Thus, each subject was confronted with only one situation, involving only one supervisor (a male for some subjects and a female for others) and one type of work group (female for some subjects, male for some, and mixed for others). From the subjects' perspective, the task was to assess the four different supervisory approaches. The experimenters' interest in the sex variable was not apparent to participants. Upon completion of the exercise, the true purpose of the study was explained, and participants were promised a summary of the findings in a future class meeting.

A  $2 \times 3 \times 2 \times 2$  analysis of variance design was used to test the effects of the experimental variables (supervisor's sex, subordinates' sex, judge's sex and judge's occupational status) on evaluations of each supervisory style.<sup>2</sup>

## RESULTS

*Effects of supervisor's sex.* Our two basic hypotheses were concerned with the effects of a supervisor's sex on people's evaluations of his or her effectiveness. First, we hypothesized that evaluations would generally be higher for male supervisors, because of the greater comparability between their supervisory role requirements and the "normal," culturally expected behavior for males. The mean ratings under the various experimental conditions are shown in Table 1. The overall

<sup>2</sup> Because of the small number of females in each occupational category, interaction effects involving the "judge's sex  $\times$  judge's status" term could not be tested. These interaction terms were pooled with the error term in testing all of the other experimental effects.

mean taken across the four supervisory styles and the three subordinate groups was 12.16 for male supervisors and 11.86 for female supervisors. Thus, the general effect is in the predicted direction, but it is not statistically significant. Therefore we do not have sufficient evidence to say that male supervisors generally are rated higher than female supervisors, when direct comparisons between the sexes are not involved.

There is stronger support for our second hypothesis—that female supervisors would be judged as more likely to succeed with certain supervisory approaches, and males with others, depending on the congruence between the supervisory approach and the culturally expected behavior for each sex. We expected that approaches involving threats and rewards would be seen as more appropriate and successful for male supervisors. The mean for the *threat* style was 6.79 for male supervisors and 6.64 for female supervisors—in the right direction but not statistically significant. For the *reward* style, the means were 11.85 for male supervisors and 10.73 for females, a significant difference ( $F = 4.36$ ,  $df = 1/239$ ,  $p < .05$ ).

We expected that the *friendly-dependent* approach would be seen as more effective for female supervisors. It was, but only to a slight degree. The male mean was 11.65 and the female mean was 11.81. However, when the sex of *subordinates* is also taken into account, an interesting phenomenon emerges. The means for the *friendly-dependent* approach, when used with subordinates of the opposite sex from the supervisor, were 12.46 for male supervisors and 12.73 for female supervisors; when this approach was used with subordinates of the same sex as the supervisor, the means were only 10.63 and 10.58, a significantly lower evaluation ( $F = 3.90$ ,  $df = 2/239$ ,  $p < .05$ ). Although we had not anticipated this phenomenon, it is consistent with our basic hypothesis regarding the influence of sex-role stereotypes. That is, both males and females probably are expected to react more favorable to intimations of dependency coming from the opposite sex.

Finally, we expected that female supervisors would be evaluated more favorably

TABLE 1  
MEAN EVALUATIONS OF SUPERVISORY STYLES

Subordinates	Male supervisors				Female supervisors				Total (n = 256)
	Male (n = 46)	Female (n = 47)	Mixed (n = 41)	All (n = 134)	Male (n = 38)	Female (n = 46)	Mixed (n = 38)	All (n = 122)	
Threat	6.82	6.78	6.77	6.79	7.39	6.82	5.70	6.64	6.72
Reward	12.06	11.36	12.14	11.85	11.52	10.20	10.56	10.73	11.31
Friendly									
Dependent	10.63	12.46	11.87	11.65	12.73	10.58	12.39	11.81	11.73
Helping	18.84	18.27	17.90	18.34	18.13	18.58	17.76	18.16	18.27
All styles				12.16				11.86	12.01

Notes. The means in this table are based on the entire sample of male and female students and bank supervisors; they are derived from individual ratings summed over the three semantic differential scales (bad-good, improper-proper, ineffective-effective). The range on each scale was 1-7. The pooled within-cell estimate of the standard deviation is 3.7.

with the *helping* approach. The means were 18.34 for male supervisors and 18.16 for female supervisors. Thus, there was a slight difference opposite to what we expected. It should be noted that this *helping* approach was evaluated extremely favorably under all experimental conditions. We can only conclude that this approach is seen as highly appropriate and congruent with cultural expectations for both males and females.

*Effects of evaluator characteristics.* We hypothesized that the sex and current occupational status of the person making the evaluations would not affect the results, since the sex-role stereotypes were assumed to be quite pervasive in our culture. The relevant statistic here was the interaction effect of evaluator's sex or occupational status and supervisor's sex. These interaction effects were not significant, thus supporting our hypothesis.

## DISCUSSION

The most interesting finding to emerge from the present investigation is that evaluations of the efficacy of certain supervisory styles are influenced by the sex of the supervisor and subordinates. A *reward* style is rated as more effective for male supervisors than for female supervisors, while a *friendly-dependent* style is rated as more effective for supervisors of either sex when used with subordinates of the opposite sex.

On the other hand, evaluations of the *threat* and *helping* styles did not differ for male and female supervisors. *Threat* was

rated extremely low and *helping* was rated high, regardless of the supervisor's sex. Thus, stereotypes of an aggressive, threatening role being appropriate for male supervisors and a compassionate, helping role being appropriate for female supervisors were not upheld by the data.

The similarity of ratings made by subjects of both sexes provides evidence that men and women share common perceptions and expectations regarding what constitutes appropriate behavior for males and females in supervisory positions. In addition, the similarity between the ratings of bankers and college students suggests that these stereotypes may be quite widely held, at least in the white-collar culture.

The relatively neutral supervisory styles employed in the present study are not necessarily a good representation of the range of behaviors falling within commonly held sex-role stereotypes for males and females. More specific types of supervisory behavior where general expectancies are clearly defined for males and females, such as highly emotional or personal behaviors, probably would heighten the observed pervasiveness of sex-role stereotypes.

In view of the unobtrusiveness of the manipulations in this experiment (subjects were unaware that the sex variable was being manipulated), these results provide clear evidence that sex-role stereotypes have an important impact on expectations regarding the appropriateness of specific supervisory behav-

iors. Many of the subjects in this study are now or soon will be subordinates and colleagues of male and female supervisors similar to those depicted in the experiment. It seems reasonable to assume that these subjects, in their occupational roles, would make their expectations known to their supervisors, thus restricting their willingness to experiment with new supervisory styles and limiting their potential effectiveness. This circular dilemma can be halted only by systematically identifying and eliminating erroneous sex-role stereotypes.

#### REFERENCES

- GILMER, B. *Industrial psychology*. New York: McGraw-Hill, 1961.  
 GILMER, B. *Industrial and organizational psychology*. New York: McGraw-Hill, 1971.

- GOLDBERG, P. Are some women prejudiced against women? *Transaction*, April 1968, 28-30.  
 HILGARD, E., & ATKINSON, R. *Introduction to psychology*. (4th ed.) New York: Harcourt, Brace & World, 1967.  
 KLEIN, X. The stereotype of femininity. *Journal of Social Issues*, 1950, 6, 3-12.  
 KREPS, J. *Sex in the market place: American women at work*. Baltimore: Johns Hopkins University Press, 1971.  
 MINER, J. *Studies in management education*. New York: Springer, 1965.  
 PIETERTSON, G., KIESLER, S., & GOLDBERG, P. Evaluation of the performance of women as a function of their sex, achievement, and personal history. *Journal of Personality and Social Psychology*, 1971, 19, 114-118.  
 SCHEINFELD, A. *Women and men*. Harcourt, Brace, 1944.  
 TYLER, L. *The psychology of human differences*. (2nd ed.) New York: Appleton-Century-Crofts, 1965.

(Received August 16, 1971)

#### Optional Blind Reviewing of Manuscripts

The *Journal of Applied Psychology* now offers its authors the option of having their manuscripts reviewed "blind" by the Consulting Editors or other reviewers. By blind reviewing it is meant that the author's identity will not be revealed to the reviewer until the latter receives copies of editorial correspondence respecting the manuscript that he has reviewed blind. Such blind reviews are subject to exactly the same standards as other submissions. The Editor will not be "blinded" at any time. To obtain a blind review: (a) Such a review must be requested of the Editor when the manuscript is first submitted. (b) The author must take sole responsibility not to reveal his identity directly or inadvertently in the manuscript itself.

## RELATIONSHIP BETWEEN MEASURES OF EFFORT AND JOB PERFORMANCE

WILLIAM E. WILLIAMS<sup>1</sup> AND DALE A. SEILER

*Western Electric Company, Princeton, New Jersey*

Effort expended and job performance are considered to be different, although not independent, constructs in the industrial environment. The relationship between these two variables was investigated using the multitrait-multimethod and multitrait-mutirater approaches. Engineer self-ratings and supervisor ratings were obtained on 202 engineers using global and dimensional rating methods. Convergent validity was found for the measures of effort and the measures of performance, but only the measures of performance demonstrated discriminant validity when compared with the measures of effort. Raters demonstrated convergent validity for each variable, but only some discriminant validity on the performance measures. Implications of the results are discussed in terms of appropriateness of the dimensional measure of effort.

Conceptually, how hard a person works (effort) is different from how well he works (proficiency). In the industrial setting, proficiency is typically considered to be synonymous with job performance, and effort can be viewed as a measure of work motivation (Landy & Guion, 1970; Porter & Lawler, 1968). Job performance, however, is not independent of effort. Porter and Lawler (1968) suggested in their model that effort leads to performance but is moderated by abilities and the degree to which the employee's behaviors are congruent with organizational goals (role perception).

Porter and Lawler (1968) and Landy and Guion (1970) have stressed the need to consider effort separate from job performance. Porter and Lawler found in their studies of managerial behavior high but far from perfect correlations between ratings of overall effort and aspects of rated performance. In addition to showing less than a perfect correlation between ratings of effort and performance, it would be desirable to demonstrate that effort and performance show discriminant validity as defined by Campbell and Fiske (1959) in their multitrait-multimethod approach. If it is postulated that there is a difference between ratings of effort and performance, this difference ideally should not

primarily be a function of either rating method or type of rater, but rather related to a difference in the variables (traits). The Campbell and Fiske methodology provides a means of assessing these relationships.

In the reported study, the Campbell and Fiske (1959) multitrait-multimethod approach was used, in part, for the purpose of exploring the relationship between the effort and performance variables ("variable" can be substituted for "trait"). Two different methods of rating the effort and performance variables (global and dimensional) and two types of raters (superior and self) were used. By making comparisons between these ratings, it was intended that a greater understanding of the relationship between effort and performance would be achieved.

### METHOD

#### Variables

Two measures of effort and two measures of job performance were obtained: (a) Landy and Guion's (1970) seven-dimension, work motivation scales, (b) Williams and Seiler's (1970) five-dimension, professional-anchored rating scales (PARS), a performance measure, (c) global measure of overall performance, and (d) global measure of overall effort.

Landy and Guion's (1970) scales (referred to as dimensional effort) were developed for engineers using the Smith and Kendall's (1963) anchored rating scale approach. For each of the seven dimensions, scaled behavioral incidents are used as reference points along a 9-point rating scale. Using the reference points as a guide, the rater selects a point on the continuum that best describes the ratee. The seven-work motivation dimensions were team atti-

<sup>1</sup> Requests for reprints should be sent to William E. Williams, Western Electric Company, Engineering Research Center, P. O. Box 900, Princeton, New Jersey 08540.

tude, task concentration, independence/self-starter, organizational identification, job curiosity, persistence, and professional identification.

Williams and Seiler's (1970) PARS performance rating scales (referred to as dimensional performance) were developed specifically for the engineer population used in the study following a procedure similar to the Smith and Kendall (1963) approach. As in the Smith and Kendall and Landy and Guion scales, scaled behavioral incidents are used as reference points for ratings on each dimension. The five behaviorally anchored job-performance dimensions were engineering proficiency, production, procedural proficiency, company identification, and administrative proficiency.

Two global ratings (referred to as global effort and global performance) were each made on a 9-point scale with the end and midpoints identified as: "a very small amount" (1), "a medium amount" (5), and "a very large amount" (9) for effort, and "very low" (1), "medium" (5), and "very high" (9) for performance. Effort was defined as how hard one works, and job performance as the overall contribution to the organization.

### Subjects and Procedures

The study was conducted in an engineering organization responsible for the development of engineering plans for the service and installation of telephone equipment. Forty-one supervisors and 202 engineers participated in the study. From 4 to 14 engineers reported to each of the supervisors.

TABLE I  
INTERCORRELATIONS AMONG VARIABLES AND  
METHODS BY ENGINEER AND  
SUPERVISOR GROUPS

Method	Variable	1	2	3
Engineers <sup>a</sup>				
Global	Effort (1)			
Dimensional	Performance (2)	.48		
	Effort (3)	.42	.67	
	Performance (4)	.38	.74	.73
Supervisors <sup>a</sup>				
Global	Effort (1)			
Dimensional	Performance (2)	.60		
	Effort (3)	.64	.83	
	Performance (4)	.59	.91	.85

<sup>a</sup> N = 202 ratings.

In small group sessions, each supervisor completed an evaluation booklet containing the two performance and two effort rating scales for each of the supervisor's engineers. The supervisor rated all his engineers on one dimension before rating them on the next dimension, etc. The order of presentation was dimensional performance, dimensional effort, global measure of performance, and global measure of effort. In small group sessions, the engineers completed a similar booklet rating themselves (self-rating). Supervisors and engineers were told the ratings were for experimental purposes, although the PARS instrument was being developed for and by the study organization to be used in their performance appraisal program. Both groups knew the other was completing the booklets.

### RESULTS

Table 1 shows the intercorrelations by supervisor and engineer groups among the effort and performance variables and the two methods. These matrices represent the multitrait-multimethod matrix used by Campbell and Fiske (1959). From these matrices the determination can be made of the convergent and discriminant validity of the variables. Convergent validity is demonstrated by a high correlation between several methods of measuring the same variable. These were the correlations between the global and dimensional ratings for effort and performance. These correlations are circled in Table 1 for the engineer and supervisory matrices. All these *r*'s are significant at the .01 level (N = 202), although the convergent correlations are much higher for performance measures than the effort measures.

Campbell and Fiske report that discriminant validity can be demonstrated in three ways. First, a variable should correlate more highly with another measure of the same variable (the circled correlations) than with any other variable having neither variable nor method in common. These latter correlations are shown in Table 1 in the squares (dotted lines). Since the circled correlations for effort (.42 for engineers and .64 for supervisors) in the two matrices fail to exceed one of the two comparison dotted line squares in each matrix, the effort measures do not show discriminant validity using the first criterion. This is not true for the performance measures where the convergent validity values (.74 for engineers and .91 for supervisors) exceeds the

comparison values in each matrix (dotted line squares). Second, a variable should correlate more highly with another measure of the same variable (the circled correlation) than with measures designed to get at different variables that happen to employ the same method. These later correlations are shown in the squares (solid lines) in both matrices in Table 1. Again, the effort measures do not show discriminant validity using the second criterion, since the effort circled correlations do not exceed the comparison dotted-line squares in each matrix. The performance measures do, however, meet the criterion.

The third way of determining discriminant validity is an examination of the similarity of patterns of correlations for submatrices within the multitrait-multimethod matrix. In a  $2 \times 2$  matrix as shown in Table 1 there are only single correlation values in the solid and dotted lined squares. In a  $3 \times 3$  or larger matrix these squares would contain additional correlations that would permit a pattern analysis within each submatrix. Since there was only a single value for these data, the third way of determining discriminant validity could not be used.

Table 2 shows the intercorrelations among the variables and raters by each rating method. This table permits a multitrait-multirater analysis as shown by Lawler (1967), in which raters are substituted for methods and the same criteria for convergent and discriminant validity are used. The convergent validity correlations are all significant at the .01 level ( $N = 202$ ), although as with multitrait-multimethod analyses, the convergent correlations are much higher for the performance measures than the effort measures. Only supervisor and engineer ratings on the performance measures met the first criterion of discriminant validity (respective circled correlations compared against the appropriate dotted-line square correlations). The supervisor and engineers ratings of effort and performance fail to meet the second criterion of discriminant validity (circled correlations do not exceed the comparison solid-lined square correlations). Meeting this last criterion is a rather stringent requirement for behavior-trait data as pointed out by Gunderson and

TABLE 2  
INTERCORRELATIONS AMONG VARIABLES AND RATERS  
BY GLOBAL AND DIMENSIONAL METHODS

Rater	Variable	1	2	3
Global ratings				
Engineer	Effort (1)			
	Performance (2)	.48		
Supervisor	Effort (3)	.24	.25	
	Performance (4)	.28	.48	.60
Dimensional ratings				
Engineer	Effort (1)			
	Performance (2)	.73		
Supervisor	Effort (3)	.33	.38	
	Performance (4)	.45	.60	.85

Nelson (1966) and Lawler (1967). The supervisor-engineer correlations come much closer, however, to meeting this criterion than the ratings on the effort variables. The third criterion for discriminant validity (pattern analysis) could not be evaluated since, as with the multitrait-multimethod matrix, there is only a single value for comparisons. Table 3 shows the intercorrelation matrix for supervisor and engineer ratings on the work motivation scales. The average correlation for supervisors was .65 and .34 for engineers (computed following a conversion of the  $r$ 's to Fisher  $z$ 's). Landy and Guion (1970) reported an intercorrelation matrix of peer ratings very similar to the reported engineers' self-ratings. The supervisor correlations show that a large halo tendency existed for their ratings.

Table 4 shows the intercorrelation matrix for supervisor and engineers ratings on the performance scales (PARS). The average correlation for supervisors was .76 and .55 for engineers (computed following a conversion of the  $r$ 's to Fisher  $z$ 's). These correlations indicate that a relatively large halo tendency existed for both ratings, although as with the work-motivation scales the supervisors exhibited greater halo.

TABLE 3  
INTERCORRELATION OF WORK MOTIVATION DIMENSIONS<sup>a</sup>

Work motivation dimension	Variable											
	1		2		3		4		5		6	
	Engineer	Supervisor										
1. Professional identification												
2. Team attitude	11	42										
3. Job curiosity	34	59	25	71								
4. Task concentration	08	51	32	59	38	69						
5. Independent/self-starter	10	48	35	70	37	79	37	71				
6. Persistence	11	46	40	67	47	75	51	70	38	74		
7. Organizational identification	29	50	39	69	34	67	39	65	39	69	47	70

Note. N = 202 ratings.

<sup>a</sup> Decimal points are removed.

## DISCUSSION

The coefficients of correlation between global measures of effort and measures of performance were similar to those reported by Porter and Lawler (1968). Porter and Lawler reported  $r$ 's of .47 and .59 for self and superior ratings, respectively. In the reported study the global effort and global performance ratings correlated .48 and .60 for self and superior ratings, respectively. Porter and Lawler concluded that such correlation coefficients were high but far from perfect, indicating that effort is a part of performance but is not the same as job performance. However, using the multitrait-multimethod approach, with the data from the reported study, measures of effort did not show discriminant va-

lidity when compared with performance measures.

The lack of discriminant validity for the effort ratings may be explained by the high correlations between the dimensional measure of effort and the measures of performance. These correlations are higher than the correlation between the two measures of effort. The correlation between dimensional effort and global performance was .67 and between dimensional effort and dimensional performance was .73. The correlation between dimensional effort and global effort was .42.

It is possible that the behaviorally anchored, work-motivation scales may not be accurate measures of effort, at least for the population of raters used in the study. Landy

TABLE 4  
INTERCORRELATIONS OF PARS DIMENSIONS<sup>a</sup>

Performance dimensions	Variable							
	1		2		3		4	
	Engineer	Supervisor	Engineer	Supervisor	Engineer	Supervisor	Engineer	Supervisor
1. Production								
2. Administrative proficiency	59	82						
3. Engineering proficiency	67	80	64	79				
4. Company identification	37	70	48	64	37	61		
5. Procedural proficiency	58	80	70	81	58	86	49	64

Note. N = 202 ratings.

<sup>a</sup> Decimal points are removed.

and Guion (1970) used peer ratings rather than superior or subordinates, which could explain the results. A good measure of effort, however, should not be rater bound. Another possible explanation for the high correlation between the dimensional effort and performance ratings was that the specific work motivation statements that anchor the rating scales were interpreted by raters as more related to the performance aspects of the job than to the effort expended aspects. In this regard a content analysis of the scales showed that some of the behavioral anchored statements used in the scales are result oriented, which may have implied to the rater the performance concept rather than simply effort expended. For example, the *organizational (company) identification* dimension appears in the PARS and work motivation scales. Specific statements related to adhering to formal and informal company policy appear in both of these scales.

The PARS performance ratings directly preceded the anchored work-motivation ratings that could have caused a response set that biased the work motivation ratings. Rather than counterbalance the order of rating (which may have been the better design, but impractical because of other considerations), very careful instructions were given for each type of rating including explanations of the variable to be rated. This does not, however, preclude the possibility of a biasing effect.

It could be argued that the performance measures used in the study were measuring effort. Although the data given cannot completely resolve the ultimate validity of the performance or effort measures, the measures of the job performance did meet the convergent and discriminant validity criteria.

Lawler (1967) showed that the job performance self-ratings by managers showed poor convergent and discriminant validity when compared with superior and peer ratings. He found correlations between self- and supervisor ratings of less than .13 ( $N = 113$ ) for performance measures. In the reported study, however, there was some convergent and discriminant validity demonstrated for supervisor and self-ratings of performance.

The correlation between self- and supervisory dimensional performance was .60 and for global performance .48 (both significant at the .01 level,  $N = 202$ ). Only one of the two discriminant validity criteria were met (see Table 4). However, as pointed out earlier, Gunderson and Nelson (1966) and Lawler (1967) indicate the criterion not met in this study is a rather stringent criterion for behavior-trait data. Supervisor and self-ratings for both effort measures did not meet the two discriminant validity criterion, although the convergent validities were significant.

Although the halo effect was not the focal point of the study, some of the dimensional ratings did show the halo effect. Supervisory ratings for both variables show high intercorrelation, although the performance intercorrelations are much higher than the effort. In addition to the general tendency for halo, the supervisors used in the study had historically made only global ratings of their engineers. It will be interesting to see if the halo effect is reduced as the supervisors gain experience in evaluating performance on a dimensional basis. The halo tendency was not as strong for engineers self-ratings, especially on the effort ratings.

The results of the study do not argue conclusively about the relationship between measures of effort and performance. Convergent validity was found within measures of effort and performance, but only discriminant validity was found for the performance measures. Similar results were found by type of rater. As noted earlier, the problem may be with the dimensional ratings of effort used. Future investigators should attempt to determine if dimensional measures of effort can discriminate between the effort and performance constructs. The overall findings would seem to indicate that in investigations aimed at separating the effort component out of performance, great care should be taken in selecting measures of both variables.

#### REFERENCES

- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.

- GUNDERSON, E. K. E., & NELSON, P. D. Criterion measures for extremely isolated groups. *Personnel Psychology*, 1966, 19, 67-80.
- LANDY, J. L., & GUION, R. M. Development of scales for the measurement of work motivation. *Organizational Behavior and Human Performance*, 1970, 5, 93-103.
- LAWLER, E. E., III. The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, 1967, 51, 369-381.
- PORTR, L. W., & LAWLER, E. E., III. Managerial attitudes and performance. Homewood, Ill.: Richard D. Irwin, 1968.
- SMITH, D. C., & KENDALL, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, 47, 149-155.
- WILLIAMS, W. E., & SEILER, D. A. Supervisor and subordinate participation in the development of behaviorally anchored rating scales. *Journal of Industrial Psychology* (in press).

(Received August 19, 1971)

Manuscripts Accepted for Publication in the  
*Journal of Applied Psychology*

- The Effect of Heredity on Attitudes toward Alcohol, Cigarettes, and Coffee. Arnon Perry (The Leon Recanati Graduate School of Business Administration, Tel-Aviv University, Tel-Aviv, Israel).
- Visual Cues and Verbal Content as Influences on Impressions Formed after Simulated Employment Interviews. Paul V. Washburn and Milton D. Hakel (Department of Psychology, College of Social and Behavioral Sciences, Ohio State University, 404-C West 17th Avenue, Columbus, Ohio 43210).
- The Impact of Performance on Managerial Pay Levels and Pay Changes. Herbert G. Heneman, III (Graduate School of Business, University of Wisconsin, 1155 Observatory Drive, Madison, Wisconsin 53706).
- Personality and Product Use Revisited: An Exploration with the Personality Research Form. Parker M. Worthington (School of Business Administration, University of Massachusetts, Amherst, Massachusetts 01002), M. Venkatesan, and Steve Smith.
- The Characteristics of Subject Matter in Different Academic Areas. Anthony Biglan (Department of Psychiatry, University of Wisconsin Medical Center, 427 Lorch Street, Madison, Wisconsin 53706).
- Relationships between Subject Matter Characteristics and the Structure and Output of University Departments. Anthony Biglan (Department of Psychiatry, University of Wisconsin Medical Center, 427 Lorch Street, Madison, Wisconsin 53706).
- The Effect of Intolerance of Ambiguity on Product Perception. Brian F. Blake (Department of Psychology, St. John's University, Grand Central and Utopia Parkways, Jamaica, New York 11432), Robert Perloff, Robert Zenhausern, and Richard Heslin.
- Organizational Independence, Learner Behavior, and Managerial Practices: A Replicated Study. Robert J. House and Steven Kerr (College of Administrative Science, Ohio State University, 1775 South College Road, Columbus, Ohio 43210).
- Race, Employment, and the Evaluation of Work. Jack Feldman (Department of Management, University of Florida, Gainesville, Florida 32601).
- The Career Choices of Married Women: Effects on Conflict, Role Behavior, and Satisfaction. Douglas T. Hall (Faculty of Administrative Studies, York University, 4700 Keele Street, Downsview 463, Ontario, Canada) and Francine E. Gordon.
- Self-Esteem as a Moderator of the Relationship between Expectancies and Job Performance. James F. Gavin (Department of Psychology, Colorado State University, Fort Collins, Colorado 80521).

## THE INFLUENCE OF VALENCE, INSTRUMENTALITY, AND EXPECTANCY ON EFFORT AND PERFORMANCE<sup>1</sup>

ROBERT D. PRITCHARD<sup>2</sup> AND MARK S. SANDERS<sup>3</sup>

*Purdue University*

An expectancy-valence model of work motivation was tested using survey methodology with a sample of government workers. The model predicted self-reported effort fairly well, but correlations with supervisory ratings of effort and performance were lower. Of the three components of the model, valence of job outcomes was by far the best single predictor. Support was given to one of the two multiplicative relationships posited by the model. Implications of the research for future testing of expectancy-valence models with survey methodology were discussed, especially for the measurement of instrumentality.

Formal expectancy-valence models of work motivation have been presented by Campbell (1969), Campbell, Dunnette, Lawler, and Weick (1970), Galbraith and Cummings (1967), Graen (1969), Porter and Lawler (1968), and Vroom (1964). While all these models offer some unique concepts, they have as a common core three basic variables: valence of job outcomes, performance-outcome instrumentality, and effort-performance expectancy. Valence of job outcomes (*V*) refers to the degree of positive or negative value, importance, or utility an individual places on intrinsic or extrinsic events that could occur on a job. Examples of job outcomes would be pay, promotion, recognition, working long hours, and feelings of accomplishment. Performance-outcome instrumentality (*I*) refers to the perceived degree of relationship a person sees between his level of performance and attaining the job outcomes. Positive values indicate that as level of performance increases, the chances for attaining the outcomes increase. For example, someone on a piece-rate payment system should have a high, positive performance-pay instrumentality since increases in performance are followed by increases in pay. Instrumentality values near zero would imply that level of performance is unrelated to attaining the out-

come, while negative values imply that the higher the performance, the lower the chances of obtaining the outcome.

Effort-performance expectancy (*E*), the third variable common to all the models, refers to the perceived degree of relationship between one's level of effort and his level of performance. High values indicate the greater the effort, the greater the performance; low values indicate that level of performance is unrelated to level of effort.

It is possible to consider both *I* and *E* to be conceptually equivalent since both refer to a perceived degree of relationship between two variables. Expectancy is the relationship between effort and performance, while instrumentality is the relationship between performance and job outcomes. This conceptual similarity presumably has led some authors (e.g., Porter & Lawler, 1968) to combine *E* and *I* into one variable and discuss the relationship between effort and job outcomes. By combining these, one has the advantage of being able to deal directly with job outcomes that are a direct function of effort. For example, it makes conceptual sense to deal with the relationship between effort and the job outcome of "feeling tired at the end of the day." It is less easy to see how this outcome could be directly related to level of performance.

While there is a conceptual advantage to combining *E* and *I* into one measure, there are advantages to keeping them separate as well. Using both variables allows one to assess the value of high performance (*V* · *I*) separately from the perceived relationship between effort and performance. In an incen-

<sup>1</sup> This research was supported by United States Postal Service contract number RER 119-70 awarded to Arthur L. Dudycha. We gratefully acknowledge this assistance.

<sup>2</sup> Requests for reprints should be sent to Robert D. Pritchard, Department of Psychology, Purdue University, Lafayette, Indiana 47907.

<sup>3</sup> Now at San Fernando Valley State College, Northridge, California.

tive pay system, for example, the value of high performance may be quite high, but due to ability, role perceptions, or external constraints, the individual may feel increased effort would not result in increased performance. In such a situation, measuring both E and I would show that the incentive system was powerful in the sense of making valued rewards contingent on performance, but that the program would not increase effort since E was low. In contrast, if one were to measure the perceived relationship between effort and job outcomes, one could not tell whether performance was seen as being unrelated to job outcomes or whether effort was seen as unrelated to performance.

Although there are these advantages to both conceptual systems, the present research deals with the original formulation of the expectancy-valence model, one in which E and I are measured separately.

Expectancy-valence models also postulate that the three components (E, I, and V) combine in a specific manner to influence effort. V and I combine multiplicatively to determine what might be called *valence of performance* ( $V \cdot I$ ). Specifically,  $V \cdot I$  equals the sum of the products obtained by multiplying the valence of each job outcome by its corresponding performance-outcome instrumentality and summing these products across all outcomes.

A second relationship considers how E and ( $V \cdot I$ ) combine to determine level of effort. Predicted level of effort is said to be the product of expectancy and valence of performance. That is, effort =  $E(V \cdot I)$ .

Research conducted to test this type of model has generally supported the model. For example, Hackman and Porter (1968), in a survey of female telephone service representatives, found that level of effort predicted by this type of model correlated .27 with supervisors' ratings of involvement and effort, and .40 with a composite effort-performance criterion. Lawler and Porter (1967) used a similar methodology with 154 managers from five organizations. They found that correlating predicted effort with supervisory, peer, and self-ratings of effort showed a median correlation of .30. Other research has also gen-

erally supported this type of model (Gavin, 1970; Georgopoulos, Mahoney, & Jones, 1957; Graen, 1969; Porter & Lawler, 1968; Shuster & Clark, 1970).

While there is some support for the prediction made by the overall model, there has been less attention given to the usefulness of the various components of the model separately. For example, the influence of E has received little attention. Tests of the model by Gavin (1970), Hackman and Porter (1968), Lawler (1968), and Porter and Lawler (1968) have combined E and I into one measure. The only study that explicitly deals with the E component is that of Graen (1969). He reported mixed findings for the relationship of this component with measures of performance.

In addition to this problem of lack of attention to the separate components of the model, a second problem deals with testing the two multiplicative relationships postulated in the models:  $V \cdot I$ , and  $E(V \cdot I)$ . While the attention given to this question (e.g., Hackman & Porter, 1968; Lawler & Porter, 1967; Porter & Lawler, 1968) has generally supported the former relationship, little attention has been given to the latter.

This study attempted to deal with these questions by (a) testing the entire model: Effort =  $E(V \cdot I)$ ; (b) measuring each component of the model (E, V, and I) separately and exploring the predictive accuracy of each component; and (c) testing both multiplicative relationships:  $V \cdot I$  and  $E(V \cdot I)$ .

## METHOD

The subjects consisted of 70 male and 76 female employees of the Post Office who were undergoing a training program to sort mail. The 30-hour training program was given over a 4- to 6-week period and consisted of memorizing a long and complex routing system. All employees were required to learn at least one of these systems. The subjects ranged in age from 18 to 45 with a median age of 22. Tenure in the Post Office ranged from 2 months to 2 years with a median of 6 months.

Interviews were conducted with agency employees and their supervisors to obtain a list of potential job outcomes. Fifteen outcomes that were mentioned often or seemed intuitively important were ultimately selected. A list of these outcomes appears in Table 1. Measures of V for each outcome were obtained through an 11-point Likert scale ranging from +5 ("Extremely good—this would be about the best

TABLE 1  
MEANS AND STANDARD DEVIATIONS OF VALENCES AND INSTRUMENTALITIES

Job outcome	Valence		Instrumentality	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1. Gaining admiration and respect from my fellow workers	2.48 <sup>a</sup>	1.94	4.51 <sup>b</sup>	3.35
2. Getting more work assigned to me	- .81	2.49	6.12	3.18
3. Being able to do my job without the help of others	2.67	2.25	7.34	2.69
4. Having my supervisor checking on my work	- .23	2.83	6.09	3.15
5. Being promoted	4.06	1.55	4.82	3.36
6. Working more consistent hours	1.86	2.85	5.53	3.21
7. Working better hours	3.87	1.88	4.67	3.34
8. Working more hours	- .26	3.04	5.23	3.28
9. Getting an opportunity to put my training knowledge to work immediately	1.97	2.74	7.51	2.65
10. Working with people who really know their jobs	3.08	1.99	6.96	2.67
11. Getting more responsible tasks to do	2.12	2.31	6.21	2.77
12. Keeping my job (i.e., not being fired)	3.86	2.19	8.12 <sup>c</sup>	3.48
13. Feeling a sense of accomplishment for mastering a difficult task	3.63	2.02	8.36	2.62
14. Being able to truly contribute to the operation of the organization	2.86	2.25	7.60	2.88
15. Getting a pay raise	4.08	1.75	4.06	3.67

<sup>a</sup> Means reflect responses to an 11-point scale ranging from -5 to +5.

<sup>b</sup> Means reflect responses to an 11-point scale ranging from 0 to 10.

<sup>c</sup> This item was worded: "The chances are \_\_\_ in 10 that learning the routing system will result in keeping my job."

thing that could happen to me on the job") through 0 ("Neutral—I don't care one way or the other whether this happens to me") to -5 ("Extremely bad—this is about the worst thing that could happen to me on the job").

The I component was measured by having respondents estimate the chances in 10 that successfully completing the training program would result in each of the job outcomes. For example, the first instrumentality item read, "The chances are \_\_\_ in 10 that learning the routing system will result in gaining the admiration and respect of my fellow workers."

The E component carries a great deal of weight in calculating predicted effort since the sum of all  $V \cdot I$  products is multiplied by the single value of E. Consequently, three items were used to measure E in hope of obtaining a more reliable measure than would be obtained using only one item. The three items were also in the "chances in 10" format and asked the probability that (a) if a person studies very hard, he will learn the system, (b) if I study very hard, I will learn the system, and (c) if a person puts in a great deal of work, effort, and home study on learning the system, he will pass the system test. The mean of these three items constituted the E measure. The median intercorrelation between the three items was .64.

A self-report measure of effort was obtained by averaging the responses to four 7-point Likert items. These items dealt with the level of effort put into learning the system, level of effort in relation to other people in the training program, level of effort in relation to the amount needed to pass the system test, and frequency of keeping up with the class

assignments. The median intercorrelation between the four items was .52.

Ratings of effort were also obtained by asking each subject's training supervisor to answer the same four items for each of his trainees. In addition, supervisors rated the performance of each trainee. Performance items included percentile performance of each trainee in comparison with all other trainees the trainer had ever dealt with, predicted ultimate efficiency at using the system, and frequency of repeating mistakes in the training program. Unfortunately, actual performance on the training program was not recorded by the trainers and thus was unavailable.

## RESULTS

Means and standard deviations of V and I for each job outcome are presented in Table 1. Inspection of the V means indicates large differences in the importance placed on different outcomes, with "promotion," "better working hours," "keeping my job," and "pay raise" being highly valued. Furthermore, there was fairly large variability across subjects within particular outcomes, especially "working more hours" (overtime) and "being able to contribute to the operation of the organization."

Analyses more directly relevant to the model are presented in Table 2. The first entry in Table 2 is the complete model,  $E(V \cdot I)$ . Some subjects did not answer every item;

TABLE 2  
COMPONENT AND CRITERION MEANS, STANDARD DEVIATIONS, AND INTERCORRELATIONS

Variables	<i>M</i>	<i>SD</i>	Intercorrelations									
			1	2	3	4	5	6	7	8	9	10
1. E(V·I)	143.51	96.67	—									
2. E	8.47	1.65	.54	—								
3. V	2.37	1.13	.86	.26	—							
4. I	6.21	1.64	.70	.32	.47	—						
5. V·I	16.23	9.92	.97	.37	.90	.71	—					
6. V·E	20.59	11.24	.94	.52	.94	.52	.90	—				
7. V + I	8.58	2.40	.88	.35	.80	.91	.91	.81	—			
8. E + (V·I)	24.70	10.65	.99	.50	.87	.71	.99	.92	.90	—		
9. E + (V + I)	17.05	3.35	.90	.74	.70	.81	.84	.83	.89	.89	—	
10. Self-report Effort	4.83	1.18	.47	.13	.54	.22	.50	.52	.41	.49	.36	—
11. Supervisory Effort	3.69	1.38	.16	.01	.22	-.02	.16	.21	.09	.15	.07	.25
12. Supervisory Performance	3.64	1.06	.17	.00	.24	.02	.17	.23	.13	.16	.09	.26

\* E = Effort-Performance Expectancy, V = Valence of Job Outcomes, and I = Performance-Outcome Instrumentality.

therefore, the sum of the  $V \cdot I$  products would, in part, be a function of the number of items answered. To eliminate this problem, valences and instrumentalities were multiplied for each outcome for which both were available, and divided by the number of complete pairs. This mean ( $V \cdot I$ ) was multiplied by E to yield the predicted effort for the entire model  $E(V \cdot I)$ . Other entries in Table 2 are also means; for example, V refers to the mean valence for all outcomes to which the subject responded.<sup>4</sup>

The complete model is a fairly good predictor of self-reported effort. However, while correlations with supervisory ratings of effort and performance are in the predicted direction, the proportion of variance accounted for by the complete model is very small.

Taking each of the three components separately, the data indicated the single best predictor is V. This component correlated with the criteria higher than did either of the other two individual components (I and E). In fact, except for self-reported effort, the other components showed no appreciable relationship with the criteria.

<sup>4</sup> It is possible that subjects who did not respond to an item actually considered it unimportant or irrelevant. To the extent this occurred, the procedure of using means would artificially inflate the size of the component for such an individual. The resulting increase in error variance would serve to attenuate the substantive relationships.

Table 2 also presents data relevant to the multiplicative relationships posited in the model. The first of these is the  $V \cdot I$  relationship. Comparison of entries 5 and 7 in Table 2 indicates that  $V + I$  resulted in lower prediction than  $V \cdot I$ . However, as will be discussed below, neither the additive nor the multiplicative combinations predicted the criteria as well as valence alone.

The second multiplicative relationship is between expectancy and valence of performance,  $E(V \cdot I)$ . Entries 1 and 8 in Table 2 show that  $E(V \cdot I)$  did not result in correlations any different from those obtained by adding the two components:  $E + (V \cdot I)$ . However, these two components correlated .99 with each other.

Table 2 also presents intercorrelations between the components in the model. It is interesting to note that while the three basic components of the model, I, V, and E, are not highly intercorrelated (median,  $r = .32$ ), the various additive and multiplicative combinations of elements are very highly correlated. In fact, the median  $r$  is .895. It is unlikely that components that are so highly intercorrelated would show strongly different relationship with the criteria.

## DISCUSSION

Taken as a whole, the data tended to offer some support for the basic expectancy-valence

model,  $E(V \cdot I)$ . The entire model correlated fairly highly with self-reported effort, but relationships with supervisory ratings were low. The multiplicative relationship between  $V \cdot I$  was supported. The multiplicative relationship  $E(V \cdot I)$  was found to predict no better than an additive relationship between the variables:  $E + (V \cdot I)$ .

There was a distinct difference in the ability of the model to predict self-reported effort as opposed to supervisory ratings of effort. It may be that some subjects were employing some sort of response set (e.g., social desirability) in completing the self-report questionnaires and tended to report high valences and large instrumentalities, as well as large amounts of effort. If other subjects were not using such a set, the resulting high correlations could have emerged due to such a response set. However, it is also possible that supervisory ratings were not especially good measures in this situation. Since much of the behavior in learning the system consisted of home study, supervisors' ratings of behavior may not reflect the total effort. It was clear that supervisors' and subjects' ratings of effort were at least *different* since the two measures correlated only .27. Consequently, it is difficult to judge whether self- or supervisory ratings are the more appropriate criterion.

One interesting aspect of the data is that the single component  $V$  (component 3, Table 2) showed higher relationships with the criteria than did any other component or combination of components. Multiplying  $V$  by  $E$  (component 6, Table 2) did not appreciably change the correlations with the criteria. However, multiplying  $V$  by  $I$  (component 5, Table 2) actually lowered the obtained correlations. One would expect that including the other components would not increase prediction if these other components had little response variability. For example, if most subjects reported the same  $I$  for a given outcome, multiplying  $V$  by  $I$  would serve to add a constant to  $V$ . Including  $I$  in such a case would neither increase nor decrease the size of the correlations. However, as Table 1 indicates,  $I$  measures actually had larger variabilities than did  $V$  measures and yet their inclusion lowered predictability. In other words, weight-

ing  $V$  by  $I$  served to increase the relative amount of error variance in the composite compared to the amount of error variance in  $V$  alone.

The error variance contributed by the  $I$  ratings may be due to the difficulty and/or ambiguity in estimating performance-outcome instrumentalities. Several things attest to this ambiguity.

First, during the administration of the questionnaires, subjects asked more questions about how to interpret the  $I$  section of the questionnaire than about any other section. One possible type of misinterpretation may have been that successfully completing the training program would result in keeping one's job, and if a person remains on the job, there is some "chance in 10" of obtaining the outcome. This can be contrasted with the correct interpretation, which was stressed to the subjects, in which the subject indicated the "chances in 10" that successfully completing the training program would *directly* result in obtaining the job outcome. If some subjects interpreted  $I$  incorrectly, this would add error variance and lower predictability.

Second, from our knowledge of the organization's functioning, the mean instrumentalities reported by the subjects appear to be overestimated in some cases. For example, completing the training program does not directly result in a promotion, yet this outcome was given a mean  $I$  of 4.82 (a probability of .482). Also, pay raises are not given after successful completion of the training program, but the reported mean  $I$  was 4.08. This seems to add support to the hypothesis that subjects may have misinterpreted the  $I$  section of the questionnaire. If so, this would account for the lowering of predictability with the inclusion of instrumentalities.

A further attempt was made to examine the possibility of an inadequate measure of  $I$ . One might assume that outcomes with low  $I$  would tend to be the most unreliably measured and thus add large quantities of error variance. If this be the case, then discounting instrumentalities with low values should increase the size of relationships with the criteria. To this end the data were reanalyzed so that all instrumentalities to which a subject

responded with a value of 4 or less were transformed to zero and the various components and intercorrelations with the criteria were recomputed.

This analysis indicated that these transformed scores resulted in correlations that were not appreciably different from the untransformed data. For example, the entire model,  $E(V \cdot I)$ , based on the transformed data, resulted in correlations of .41, .18, and .18 with self-ratings of effort, supervisors' ratings of effort, and supervisors' ratings of performance, respectively. For the original, untransformed data the corresponding correlations were .47, .16, and .17. Thus, the sources of error variance present in the I measure are not confined to low values of I.

Two major implications for tests of expectancy-valence models with survey methodology emerged from this study. First, one should insure that variability in I (and E) actually does exist in the sample. Unless this condition is met, these components cannot add to prediction. However, as long as the survey is limited to one organization, one would expect, assuming accurate perceptions, that I's would be relatively constant. For example, the organization may very well tend to promote, give pay raises, etc., on the same basis throughout the organization. Thus, actual I's may be fairly constant for all the subjects in the sample and not add much to prediction. One way to minimize this problem would be to draw a sample from different organizations with different promotion policies, pay policies, chances for feelings of accomplishment, etc. This would maximize the chances for variability in I.

A second implication from this study is the great care that should be taken in measuring the components of the model, especially the I component. This is especially true since the components are multiplied. Even a good measure of valence if multiplied by instrumentalili-

ties with large chunks of error variance will undoubtedly result in low relationships with behavior.

#### REFERENCES

- CAMPBELL, J. C. Managerial motivation: An overview. Paper presented at the Midwestern Psychological Association meeting, Chicago, May 1968.
- CAMPBELL, J. C., DUNNETTE, M. D., LAWLER, E. E., III, & WEICK, K. E. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.
- GALBRAITH, J., & CUMMINGS, L. L. An empirical investigation of the motivational determinants of task performance: Interactive effects between instrumentality-valence and motivation-ability. *Organizational Behavior and Human Performance*, 1967, 2, 237-257.
- GAVIN, J. F. Ability, effort, and role perceptions as antecedents of job performance. *JSAS, Catalog of Selected Documents on Psychology*, 1970, Issue 5, 1-26.
- GEORGOPOLOUS, B. S., MAHONEY, G. M., & JONES, N. W. A path-goal approach to productivity. *Journal of Applied Psychology*, 1957, 41, 345-353.
- GRAEN, G. B. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology*, 1969, 53, 1-25.
- HACKMAN, J. R., & PORTER, L. W. Expectancy theory predictions of work effectiveness. *Organizational Behavior and Human Performance*, 1968, 3, 417-426.
- LAWLER, E. E., III. A correlational-causal analysis of the relationship between expectancy attitudes and job performance. *Journal of Applied Psychology*, 1968, 52, 462-468.
- LAWLER, E. E., III, & PORTER, L. W. Antecedent attitudes of effective managerial performance. *Organizational Behavior and Human Performance*, 1967, 2, 122-142.
- PORTER, L. W., & LAWLER, E. E., III. *Managerial attitudes and performance*. Homewood, Ill.: Irwin-Dorsey, 1968.
- SHUSTER, J. R., & CLARK, B. Reviewing portions of the Porter and Lawler theoretical model. *JSAS, Catalog of Selected Documents in Psychology*, 1970, Issue 7, 1-19.
- VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

(Received July 27, 1971)

## EXPECTANCY THEORY PREDICTIONS OF ACADEMIC EFFORT AND PERFORMANCE<sup>1</sup>

TERENCE R. MITCHELL<sup>2</sup> AND DELBERT M. NEBEKER

*University of Washington*

Expectancy theory models were used to predict the effort and performance of college students. The expectancy theory suggests that effort is related to the degree to which the behavior (or job) is seen as leading to various outcomes weighted (multiplicatively) by the evaluation of these outcomes. This model was supported by the data reported here. The predictability of effort was increased by including extensions of the effort model by adding others' expectations and perceived influence. The job performance model suggests that effort and ability combine to predict performance. Neither the additive nor the multiplicative models found support in this setting. The extensions and modifications of the theory are discussed in detail.

Over the last 40 years, numerous psychologists have argued that an individual's behavior is a function of the degree to which the behavior is instrumental for the attainment of some outcomes and the evaluation of these outcomes (see Mitchell & Biglan, 1971, for a review). For the purposes of clarity, we will refer to the theory as expectancy theory, although other names such as instrumentality theory or social learning theory have also been used. The following research was designed to test the ability of this theory (and some recent modifications) to predict the effort and performance of college students.

A review of the research in this area indicated that very little had been done with expectancy theory to predict academic success. Todd, Terrell, and Frank (1962) report that students who believed that their endeavors were likely to lead to academic success were more likely to be normal achievers than underachievers. Battle (1965) also found that persistence on academic tasks was related positively to the expectancy of successful accomplishment. She reports a correlation of .47 ( $p < .001$ ) between the expected grade in mathematics and the persistence (time spent) of seventh- through ninth-grade children working on math problems.

<sup>1</sup>This study was supported in part by Contract NR177-472, N00014-67-A-0103-0013, Office of Naval Research, Department of the Navy (Fred E. Fiedler, Principal Investigator).

<sup>2</sup>Requests for reprints should be sent to Terence R. Mitchell, Organizational Research, University of Washington, 33 Johnson Hall, Seattle, Washington 98105.

The results to date, therefore, appear rather supportive. However, recent reviews and modifications of the theory have suggested clarifications and additions that supposedly should increase the theory's predictability (see Campbell, Dunnette, Lawler, & Weick, 1970; Mitchell & Biglan, 1971; Vroom, 1964). A brief review of these changes and their implications for the present research follows.

### THEORETICAL MODELS

#### *Job Effort*

The job effort model contends that one exerts a certain amount of effort based on three factors: (a) the degree to which effort is seen as leading to good performance, (b) the degree to which good performance is instrumental for the attainment of some outcomes, and (c) the evaluation of these outcomes.

Symbolically,  $W = E(\sum_{i=1}^N I_i V_i)$  where

$W$  = amount of work (effort),

$E$  = expectancy, i.e., the degree to which effort leads to successful performance,

$I_i$  = the instrumentality of performance for the attainment of the  $i$ th outcome,

$V_i$  = the valence or importance of the  $i$ th outcome, and

$N$  = the number of outcomes.

Thus, one works hard if (a) he thinks his effort will lead to good performance ( $E$ ) and (b) he believes that good performance will lead to valued outcomes ( $\sum_{i=1}^N I_i V_i$ ).

Four major modifications of this model will be tested in the following study. First, Fishbein (1965) and Rosenberg (1956) have data that support the idea that one's attitude about an

object or a behavior is equal to the degree to which that object or behavior is linked to other outcomes multiplied by the evaluation of those outcomes. Fishbein (1965), for example, reports correlations of .80 between a direct measure of attitude (four bipolar scales) and a measure of the  $\sum IV$ . Given this relationship, one might argue that a direct assessment of performance ( $A_p$ ) using bipolar scales would be more parsimonious than the measurement of a whole set of instrumentalities and valences as demanded by the model above. Fishbein (1967) has also argued that one's attitude toward an act is equal to the degree to which the behavior is linked to valued outcomes. Therefore, a direct attitude assessment of effort ( $A_w$ ) should equal  $E(\sum IV)$ . By gathering both the components of the model and the direct bipolar assessment of effort and performance, we should be able to test these predictions.

A second modification suggested by Dulany (1968), Fishbein (1967), and Graen (1969) is that additional components should be added to the theory. They argue that our behavior is also determined by the surrounding social environment. Thus, we behave in a certain way not only because we believe it will lead to certain payoffs but also because we wish to fulfill the expectations of those around us. Therefore, effort will be predicted both with and without the inclusion of an expectation measure (e.g., to what extent do your peers expect you to spend time on academic activities?). The revised model is  $W = E(\sum_{i=1}^N I_i V_i) + E_p + E_f$  where

$E_p$  = expectations of peers and

$E_f$  = expectations of faculty.

One could include "fulfillment of expectations" as an outcome within the  $\sum IV$ . However, the models presented by Dulany (1968), Fishbein (1967), and Graen (1968) treat these variables separately. A third suggestion comes from the research of Dulany (1968), Fishbein (1967), and the review by Mitchell and Biglan (1971). These authors argue that this theory is essentially predicting an *intention* to behave in a certain way. So, for example, based on an expectancy theory equation, we might predict that a given student was going to spend the evening studying. However, a number of things might stop him from carrying out that

intention such as a flat tire on the way to the library or the fact that the books he needed were already checked out. The degree to which one can carry out his intentions is due partially to the degree to which one has control over the behavior in question. Therefore, we would predict that the equations given above will do a better job of predicting effort for those students who indicate a high degree of control over their academic behavior than for those who say they lack this control.

A final suggestion is that the outcomes be split into separate categories. A number of authors have presented data indicating that intrinsic factors are better motivators than extrinsic ones (Campbell, et al., 1971; Mitchell & Albright, 1971). Therefore, it is hypothesized that one's effort may be related more to the degree to which intrinsic outcomes are obtained than to the degree to which extrinsic outcomes are obtained.

#### Job Performance

The performance models presented by Vroom (1964) and Porter and Lawler (1968) have postulated that performance can be predicted by an effort  $\times$  ability score. The expectancy equation is essentially the motivational or effort component. However, in most of the reported research (Graen, 1969; Hackman & Porter, 1968; Lawler, 1968; Porter & Lawler, 1968) performance is predicted from the motivational component without the use of an ability measure. In the one investigation using an ability measure (Arvey & Dunnette, 1970) the authors report that ability was significantly related to performance, but the ability  $\times$  expectancy measure was not. A multiple correlation coefficient using ability, expectancies, and their interaction as separate predictors (i.e.,  $Performance = Ability + Expectancy + Ability \times Expectancy$ ) was significant. They argue that, perhaps, an additive relationship between ability and expectancy is a better predictor of performance than a multiplicative one.

Other authors have debated this issue (see Vroom, 1964, for a review). Therefore, it was decided to test both an additive and multiplicative model in the following investigation. Measures of ability will be added to and multiplied by the motivation component [ $E(\sum IV)$ ] to predict performance.

A summary of the suggested hypotheses are listed below.

1. Effort (W) can be predicted from the equation  $W = E(\Sigma IV)$ .
2. Effort (W) can be predicted better from a direct attitude measure ( $A_p$ ) than from the  $\Sigma IV$  for performance.
3. Effort (W) can be predicted better from a direct attitude measure of effort ( $A_w$ ) than from the whole equation [ $E(\Sigma IV)$ ].
4. Effort (W) can be predicted better when the expectations of those around the individual ( $E_p + E_t$ ) are included as predictors than when they are omitted.
5. Effort (W) can be predicted better for those individuals who feel they have control over their behavior than for those who don't have this opinion.
6. Effort (W) can be predicted better from intrinsic outcomes than from extrinsic ones.
7. Performance (P) can be predicted better from a multiplicative relationship between effort (W) and ability (A) than from an additive relationship.

## METHOD

### Subjects

Sixty male undergraduates from the University of Washington participated in the experiment. Participation was voluntary and subjects were assured that all the information given would be made public only in summary form. Nine or 10 subjects, depending on the analysis, were dropped because of missing data.

*Performance (P).* The subjects' grade point averages (GPA) for the last quarter were obtained (with their permission) from the academic files.

*Ability (A).* Upon entering the university, each subject took a battery of tests known as the Washington Pre-College Entrance Exam. Scores from these tests were combined with other data (e.g., high school average), and a predicted GPA is generated by means of a multiple linear regression equation. It was this predicted grade point average that was used as an ability measure; again, with the subjects' permission.

*Effort (W).* The subject indicated the average number of hours per week spent on academic activities for the last quarter.

### Job Effort Model Measures

*Outcomes.* The selection of outcomes for this study was based on two factors. First, outcomes were solicited from 10 students, and the final list represented almost all of their suggestions. Second, the outcomes chosen appeared to represent those outcomes that were most strongly related to satisfaction in the Constantinople (1967) study where expectancy theory was used. Nine outcomes were chosen. Three were considered to be intrinsic—feelings of accomplishment, self-confidence, and appreciation of ideas. Two were labeled extrinsic and impersonal—a good job and admission to graduate school. Four were classified as extrinsic and social, socially attractive (other sex and same sex), parental praise, and respect from peers.

*Valence (V).* The nine outcomes were listed with the letters a-i to their left. Subjects rated the degree to which obtaining or maintaining a high level of each

outcome was important and pleasant. The response was indicated by placing the letter corresponding to each outcome in the appropriate box on two 7-point scales with values ranging from +3 to -3. More than one letter could be placed in each box. The valence estimate was the mean of these two scores.

*Expectancy (E).* The subject estimated on a 7-point scale the degree to which he felt that the time he spent on academic activities would lead to good grades. Scale values ranged from 0 to 6.

*Instrumentality (I).* Since the measure of instrumentality reflects the relationship between *performance* and the *outcomes*, an estimate was made by the subject as to the degree to which obtaining good grades contributed to or detracted from the possibility of obtaining each outcome. The rating was made on a 7-point scale with values ranging from +3 to -3. More than one letter could be placed in any box.

*Attitude toward effort (A<sub>w</sub>).* Ratings of the pleasantness and importance of the time spent on academic activities was made on two bipolar scales. The mean of these ratings was used as an estimate of  $A_w$ .

*Attitude toward performance (A<sub>p</sub>).* The mean ratings of scales assessing the pleasant-unpleasant and important-unimportant feelings about good grades was used as the estimate of  $A_p$ .

*Expectations (E<sub>p</sub> + E<sub>t</sub>).* Subjects indicated the amount of time their peers (i.e., students with whom they spent most of their time last quarter) and their professors expected them to spend on academic activities. Seven-point bipolar scales going from "a great deal of time" to "very little time" were used.

*Control (C).* The amount of control that the subject felt he had over the amount of time he spent on academic activities was rated on a complete control to no control 7-point bipolar scale.

On each scale where subjects made more than one response to a given question (e.g., instrumentality), these scores were standardized around the subject's own mean. This procedure should lessen the effects of response sets that would increase the effects of measurement error when correlating across subjects. See Mitchell (1971) for a further discussion of this point.

## Job Performance Model

This model postulates that performance can be predicted from estimates of effort and ability. Performance (P) and ability (A) were defined in the criteria section. Three estimates of effort were used. The first is the time spent, which we have labeled W. The second is attitude toward effort ( $A_w$ ) and the third is our motivational [ $(E(\Sigma IV))$ ] and expectations model ( $E_p + E_t$ ).

## RESULTS

### Job Effort Model

Hypotheses 1 through 4 dealt with the job effort model and some extensions of the model. These extensions were concerned with substituting attitude measures for the motivational components of the theory and the addi-

TABLE 1

PREDICTION OF EFFORT (W) FROM THE JOB EFFORT MODEL AND ADDITIONAL COMPONENTS

	Predictor	Correlation coefficient
Motivational estimates	$A_w$	.27*
	$E(A_p)$	.33*
	$E(\Sigma IV)$	.23*
Expectations	$E_p$	.49**
	$E_f$	-.05
	Multiple $R$	
	$A_w + E_p + E_f$	.53**
	$E(A_p) + E_p + E_f$	.55**
	$E(\Sigma IV) + E_p + E_f$	.51**

Note.— $N = 51$ .

\*  $p < .05$ .

\*\*  $p < .01$ .

tion of social components dealing with the expectations of others. Three motivational measures were used:  $A_w$  = the attitude towards effort;  $E(A_p)$  = the expectancy that effort leads to good grades weighted by the attitude toward good grades;  $E(\Sigma IV)$  = the expectancy that effort leads to good grades multiplied by the sum of good grades leading to outcomes weighted by the importance of the outcomes. Table 1 presents the relevant coefficients.

These results provide relatively good support for the job effort model. Support for the substitution of attitude measures for the total  $E(\Sigma IV)$  score (i.e.,  $A_w$ ) or for just the  $\Sigma IV$  score (i.e.,  $A_p$ ) could come from two sources. First, are they related to other measures in the way that the theory suggests they should be? The somewhat similar amounts of predictability suggest that they are interchangeable. Parsimony would demand the use of the simpler measure. However, a second source of support was questionable. The intercorrelations of the three measures were .52 for  $A_w$  and  $E(A_p)$ , .37 for  $A_w$  and  $E(\Sigma IV)$ , and .71 for  $E(A_p)$  and  $E(\Sigma IV)$ . Although all three of these coefficients are significant ( $p < .01$ ), they provide only moderate support for the idea that they are measuring the same construct. The use of a direct attitude measure would also mean that the instrumentality and valence measures would be omitted. This

information can be useful in understanding our results in later analyses, and it is, therefore, suggested that the use of the attitude measure instead of gathering the additional  $E(\Sigma IV)$  is probably not a good idea.

The second modification—the addition of expectations—was strongly supported. The increased predictability, however, seems to be added by peers' expectations, with faculty expectations accounting for essentially none of the variance of effort scores.

In summary, then, the job effort model received good support with both motivational and expectation components controlling significant amounts of variance of the effort estimates.

Hypothesis 5 suggested that we would obtain better prediction of effort for subjects who indicated that they had control over their time spent than for subjects who indicated that they had little control over their effort. Subjects were split at the median, and Table 2 presents the data illustrating this hypothesis.

Two inferences from this Table should be discussed. First, the overall predictability for high-control subjects is slightly less than for low-control subjects. This result is rather

TABLE 2

PREDICTION OF EFFORT FOR SUBJECTS WHO INDICATED HIGH/LOW CONTROL OVER THEIR TIME SPENT

	Predictor	Correlation coefficients	
		High-control subjects	Low-control subjects
Motivational estimates	$A_w$	.31*	.16
	$E(A_p)$	.41**	.19
	$E(\Sigma IV)$	.25*	.13
Expectations	$E_p$	.52**	.48**
	$E_f$	.03	-.18
	Multiple $R$		
	$A_w + E_p + E_f$	.54**	.56**
	$E(A_p) + E_p + E_f$	.55**	.62**
	$E(\Sigma IV) + E_p + E_f$	.53**	.57**

Note.— $n = 25$  for high-control and 25 for low-control subjects.

\*  $p < .05$ .

\*\*  $p < .01$ .

difficult to interpret in that zero order correlations between four of the five predictors and effort is lower for low-control than for high-control subjects.

The second inference, however, is that the motivational components are clearly more related to effort for high-control subjects than for low-control subjects. Overall, the results make sense using the following rationale: Those who indicate little control over how much time they spend should be more influenced by others. Indeed, it appears that the predictability for low-control subjects is more related to others' expectations than for high-control subjects. Those who have high control, however, seem to have their effort systematically related to what they believe to be the consequences of their effort. Thus, the use of a control measure influences the relationship between our motivational and expectation measures and effort in different ways. The more control, the more the subjects' effort appears to be instrumental. The less control, the more the subjects' effort is influenced by the expectations of others.

The last hypothesis for the job effort model (number 6) dealt with the contributions made to the motivational component [ $E(\sum IV)$ ] by the instrumentalities and valences of our three sets of outcomes. The subjects were divided into two groups at the median effort score, and an analysis of variance was performed on their  $\sum IV$  scores for the social and extrinsic impersonal factors. There was a main effect for both effort ( $F = 6.03, p < .05$ ) and outcomes ( $F = 35.23, p < .01$ ) with a nonsignificant interaction.

The main effect for effort simply reflects higher  $\sum VI$  scores for high-effort subjects. This relationship was also inferred from our significant correlations between our estimates of effort ( $W$ ) and the  $E(\sum IV)$ . The interesting result of the analysis was that the contributions made by the intrinsic and extrinsic impersonal outcomes ( $\bar{X} = 3.03, 4.05$  respectively) to the total  $\sum IV$  was much greater than the contribution made by the extrinsic social factors. ( $\bar{X} = 1.36$ ) Although we hypothesized that the intrinsic components would be important, we did not do so for the extrinsic impersonal components.

Our final analysis attempted to break the

TABLE 3  
ANALYSIS OF VARIANCE FOR VALENCES AND INSTRUMENTALITIES

Factor	Anova for valences		
	Intrinsic	Outcomes	
		Extrinsic Social	Extrinsic Impersonal
Effort	High ( $n = 24$ )	2.67	2.03
	Low ( $n = 26$ )	2.53	1.90
Test	$F$	Sign	
Effort	3.21	$p < .10$	
Outcomes	28.58	$p < .01$	
Interaction	1.48	$ns$	

Anova for instrumentalities				
Factor	High ( $n = 24$ )	1.55	1.11	2.44
Effort	Low ( $n = 26$ )	1.13	.91	2.43
Test	$F$	Sign		
Effort	1.71	$ns$		
Outcomes	104.93	$p < .01$		
Interaction	1.89	$ns$		

$\Sigma IV$  down into its component parts (valences and instrumentalities). Table 3 presents these data. Here, we find that it is the intrinsic factors that are most valued and the extrinsic impersonal outcomes perceived as most attainable. This explains why the  $\Sigma IVs$  for the intrinsic and extrinsic impersonal factors were higher than the  $\Sigma IV$  score for the extrinsic social ones. It also suggests that there is a mismatch between what is valued and what is obtained; students perceive good grades as instrumental for obtaining outcomes that are not their most highly valued outcomes.

#### Job Performance Model

The job performance model suggests that estimates of effort combine with ability to predict performance. Both multiplicative and additive combinations of these variables were examined. The two models were compared using two types of mathematical models. The first was a simple combination of the standardized measures of effort and ability by either addition or multiplication and, then,

TABLE 4  
PREDICTION OF GRADE POINT AVERAGE FROM ADDITIVE AND MULTIPLICATIVE MODELS

Predictors		Simple combinations correlation coefficients		Multiple regression combinations multiple <i>R</i> s	
		Additive	Multiplicative	Additive	Multiplicative
Ability + Effort	A; W	.45*	.42*	.58*	.60*
Ability + Motivational estimates	A; A <sub>w</sub> A; E(A <sub>p</sub> ) A; E( $\Sigma$ VI)	.57* .55* .46*	.56* .53* .47*	.60* .62* .59*	.63* .64* .61*
Ability + Motivational estimates + Expectations	A; (A <sub>w</sub> + E <sub>p</sub> + E <sub>f</sub> ) A; (E(A <sub>p</sub> ) + E <sub>p</sub> + E <sub>f</sub> ) A; (E( $\Sigma$ IV) + E <sub>p</sub> = E <sub>f</sub> )	.41* .41* .37*	.41* .42* .37*	.57* .57* .57*	.59* .60* .60*

Note.—For the combinations using all four predictors, the ability measure was combined with the  $\Gamma'$  estimate of effort generated from the multiple regression equations used to predict effort from motivation and expectation.

*N* = 50.

\*  $p < .01$ .

correlating the result with performance. The second was a multiple regression approach that made a least squares fit of ability and effort with performance. As a test of the additive model, this method is clear but as a test of the multiplicative model, it was necessary to make a log transformation of the dependent and independent variables since  $\log P = \log A + \log W$  is equal to  $P = AXW$ . Table 4 presents the data for both models and methods using our measure of effort (W), our three measures of motivation, and the extended model using both motivational and expectation components. The zero order correlations between these components indicate that only ability was strongly related to performance ( $r = .57$ ,  $p < .01$ ) and hence is responsible for the models predictions.

These data seem to support the following conclusions. First, no difference can be found between the additive and multiplicative models in predicting performance in this situation. All of the possible comparisons shown in Table 4 support this inference. Second, our measures of effort are not related to performance in the academic setting. Third, since measures of ability may already include elements of effort, and our measures of effort were not found to be strongly related to performance, the present job performance models may be inappropriate for the academic setting. Fourth, the multiple

regression method cannot be said to be superior to the simple combination method even though it has larger values because the multiple regression solution is a least squares "fit" to the data while the simple combination is not.

#### SUMMARY AND CONCLUSIONS

Our results for the job effort model were generally supportive. The use of attitude measures to estimate components of the model fit into the theory as they should have. We suggested, however, that substituting the attitude measure for the other components would mean losing information about expectancies, instrumentalities, and valences. This latter information proved to be useful in understanding in more detail the relationship between effort and the components of the model. For example, the results showing that students see their effort as most instrumental for attaining impersonal goals while they value intrinsic ones most suggests that there is a mismatch between what students would like to get out of college (and what educators would like to provide) and what they believe they will obtain with good grades.

It was also clear that the inclusion of others' expectations contributed significantly to the prediction of effort. An interesting result was that while student expectations were related positively to effort, the expectations of faculty

were unrelated to student effort. These data provide little support for the idea that high expectations on the part of a teacher will be highly motivating for the students.

A third finding of interest was that the perceived control on the part of subjects was related differentially to the components of the model and their relationship to effort. More specifically, for subjects who felt they had high control, the expectations of others was less related to effort than when this perceived control was low. Intuitively, this makes sense. The less control we have, the more we are influenced by others.

Our tests of the job performance model seem to indicate that there is no difference between additive and multiplicative combinations of effort and ability to predict performance. The lack of difference between the additive and multiplicative models when simply combined also indicates that effort is not interacting with ability to mask its effect. An interesting result was that effort ( $W$ ) was unrelated to GPA while the motivational components ( $A_w$ ;  $E(A_p)$  and  $E(\sum IV)$ ) only controlled small amounts of variance in GPA. These data suggest that in this situation, neither the students' hours spent or his motivation to work are related to his grades, while ability (predicted GPA) was strongly related to grades. We suspect that the contributions made by these variables will be different in other situations. It is, however, an interesting comment on what it takes to be a successful student.

In summary, the current investigation provided some support for expectancy theory predictions of effort and performance. Some theoretical modifications were suggested and supported, and a number of interesting and substantive findings were reported as subject areas for future investigations.

#### REFERENCES

- ARVEY, R. D., & DUNNETTE, M. D. Task performance as a function of perceived effort-performance and performance-reward contingencies. (Center for the Study of Organizational Performance and Human Effectiveness Tech. Rep. No. NR 152-293) Minneapolis, Minn.: University of Minnesota, 1970.
- BATTLE, E. S. Motivational determinants of academic task persistence. *Journal of Personality and Social Psychology*, 1965, 2, 209-218.
- CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial behavior, performance and effectiveness*. New York: McGraw-Hill, 1970.
- CONSTANTINOPLE, A. Perceived instrumentality of the college as a measure of attitudes toward college. *Journal of Personality and Social Psychology*, 1967, 5, 196-201.
- DULANY, D. E. Awareness, rules, and propositional control: A confrontation with S-R behavior theory. In D. Horton & T. Dixon (Eds.), *Verbal behavior and general behavior theory*. Englewood Cliffs, N.J.: Prentice-Hall, 1968.
- FISHBEIN, M. A consideration of beliefs, attitudes, and their relationships. In I. D. Steiner & M. A. Fishbein (Eds.), *Current studies in social psychology*. New York: Holt, Rinehart & Winston, 1965.
- FISHBEIN, M. Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement*. New York: Wiley, 1967.
- GRAEN, G. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology* 1969, 53 (2, Pt. 2).
- HACKMAN, J. R., & PORTER, L. W. Expectancy theory predictions of work effectiveness. *Organizational Behavior and Human Performance*, 1968, 3, 417-426.
- LAWLER, E. E. A correlational-causal analysis of the relationship between expectancy attitudes and job performance. *Journal of Applied Psychology*, 1968, 52, 462-468.
- MITCHELL, T. R. Instrumentality theories: Conceptual and methodological problems. *JSAS Catalog of Selected Documents in Psychology*, 1972, 2, 37.
- MITCHELL, T. R., & ALBRIGHT, D. Expectancy theory predictions of the satisfaction, effort, the performance, and retention of naval aviation officers. *Organizational Behavior and Human Performance*, 1972, 8, 1-20.
- MITCHELL, T. R., & BIGLAN, A. Instrumentality theories: Current uses in psychology. *Psychological Bulletin* 76, 432-454.
- PORTER, L. W., & LAWLER, E. E. *Managerial attitudes and performance*. Homewood, Ill.: Erwin-Dorsey, 1968.
- ROSENBERG, J. Cognitive structure and attitudinal affect. *Journal of Abnormal and Social Psychology*, 1956, 53, 367-372.
- TODD, F. J., TERRELL, G., & FRANK, C. E. Differences between normal and underachievers of superior ability. *Journal of Applied Psychology*, 1962, 46, 183-190.
- VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

(Received August 17, 1971)

## SAFETY TRAINING BY ACCIDENT SIMULATION<sup>1</sup>

STANLEY RUBINSKY<sup>2</sup> AND NELSON SMITH

*University of Rhode Island*

A bench grinder was modified to allow the simulation of an accident when an unsafe operation was performed. Accident simulation was used as a training technique and compared with training by the use of written instructions and demonstrations. Subjects trained by accident simulation methods performed significantly fewer unsafe acts and retained their superior habit pattern for at least six months. Further, it was found that the training effect was transferable to a similar but unmodified tool. The use of accident simulation holds promise as a powerful and effective training technique.

Traditionally, instruction in the safe operation of power tools and machines relies on verbal and written instructions and, to some extent, on demonstrations of safe operating procedures. It is reported (U.S. Department of Labor, 1971) that the 1969 injury rate in industry was the highest since 1951 and has continued to worsen since 1958. Clearly, innovative techniques for teaching safe operation of power tools and machines are badly needed.

One training approach that has received insufficient attention in industry is "accident stimulation." While a number of earlier writers (e.g., Heinrich, 1950; Vaughn, 1928) strongly implied that actual "experience" of an accident should have a marked effect on subsequent behavior, little empirical data on this is available. More recently, Gibson (1961) called attention to the need for devices that simulate particular dangers while allowing for subjects to act (safely or unsafely). The limitations of such simulations were discussed by Haddon, Suchman, and Klein (1964); these included the possibility of injury to the subject, certain ethical issues, and the artificiality of studying accidents outside the environments in which they occur. Recently, Rubinsky and Smith (1970) developed a device that attempted to meet the objectives cited by Gibson, while minimizing the problems outlined by Haddon, et al. The device of

Rubinsky and Smith simulated accidents that could occur with the improper use of an off-hand grinder, a common industrial machine.

A source of possible injury inherent in the use of an off-hand grinder is the explosion of the grinding wheel. An unsuitable, unbalanced, or a cracked wheel will explode when subjected to the inertial forces imposed as it accelerates to its high rotational speed during normal operation. To avoid injury in the event of an exploding wheel, it is necessary only that the operator stand to the side of the machine, out of the plane of rotation of the grinding wheel during the startup phase of the grinding operation. When the wheel has attained its full operating speed, most of the danger of an exploding wheel has passed.

This article describes three experiments using the device which simulates accidents involving such an exploding bench grinder wheel. Specifically, the use of accident simulations as a training method was compared to training by written instructions and demonstrations.

### EXPERIMENT I

In this experiment, three different methods of the presentation of the simulated accident and their effect on retention of training for a 1-week period were investigated. In addition, the retention test was conducted on a different but similar grinder located in a different laboratory.

### METHOD

#### *Subjects*

The subjects were 32 male college sophomores who volunteered from an introductory psychology class. They were randomly divided into four groups.

<sup>1</sup> The data on which this article is based was performed pursuant to Contract No. PH 86-68-207 with the United States Public Health Service, Department of Health, Education, and Welfare.

<sup>2</sup> Requests for reprints should be sent to Stanley Rubinsky, Department of Industrial Engineering, University of Rhode Island, Gilbreth Hall, Kingston, Rhode Island 02881.

### Apparatus

The simulator was a standard bench grinder with two water jets attached. When a switch was closed by the experimenter, a spray of water was directed at the operator's normal position in front of the grinder; this was the simulated accident. Thus, when an accident was simulated, an operator standing in front of the machine was "injured" while an operator standing in the correct position to the side of the grinder was not sprayed with water and, therefore, escaped "injury."

In addition, 10 steel rods, numbered from 1 to 10, were used for a spark test. Each of the 10 rods was of a different composition so that their spark characteristics, while being ground, would vary. A chart, exhibiting the spark patterns of various steels was displayed directly behind the grinder as a guide to aid the subject in identifying the type of steel in the individual rods. Finally, a special tool rest with a "V" notch to accommodate the steel rods was attached to the grinder and safety goggles were provided.

### Procedure

A cover task was used so that the subjects would be unaware of the true nature of the experiment and not concentrate exclusively on the safe operation of the grinder. The subjects were told that they were participating in an experiment to see if inexperienced people could use a grinder to identify the chemical compositions of various bars of metal by their sparking characteristics when held against the grinding wheel.

**Training.** Each of the four groups was administered a different type of safety training. Group 1 was given written instructions (including safety procedures) for a grinding task; in addition, the task, including the safe operation procedures, was demonstrated. This group was the control and represented a safety training method used in industry. Group 2 was instructed in precisely the same manner except that the simulated accident, consisting of the investigator turning on the water jets was shown and explained during the demonstration. Instructions for Group 3 duplicated those for the control group except that the simulated accident was omitted from the demonstration. Instead, each subject in the group was subjected to a simulated accident during a trial run that was part of the instructions to this group. Group 4 received the same basic instructions as had all previous groups but in addition, the simulated accident was both demonstrated to and experienced by each subject.

All subjects then tested each of the 10 steel bars. They turned the grinder off between each test while they entered their judgment of the type of steel ground on a data sheet at the investigator's desk. Thus, there were 10 opportunities for "accidents" to occur. An "accident" was scored if the subject was standing in front of the grinding during the startup period.

One week later, the subjects returned to a dis-

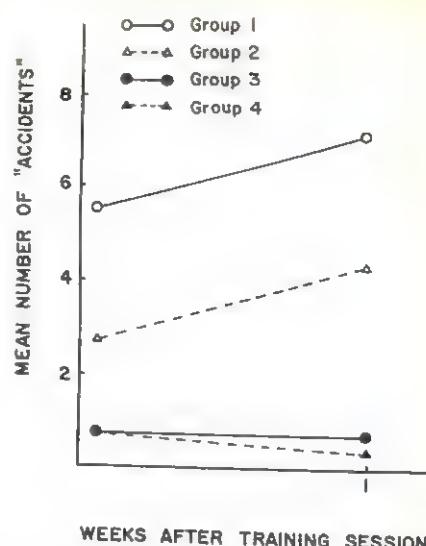


FIG. 1. Mean number of "accidents" for subjects within training conditions during training and one retention session ( $N = 32$ ).

ferent laboratory and again tested each of the steel bars. The safety instructions were not repeated for this test.

### RESULTS

The mean number of accidents occurring during the training and replication sessions are shown in Figure 1. It can be seen that the mean number of accidents decrease from Group 1 to Group 4.

An analysis of variance showed that the differences between groups, sessions, and their interactions were all significant ( $F = 13.68, 8.28, 5.33; 3/28, 1/28, 3/28 df; p < .001$ ).

To further identify the effect of the different training procedures, a series of  $t$  tests was performed between the groups for both training and replication. The results of this test for the training session showed that there were significantly fewer "accidents" in all experimental groups compared to the control group. The  $t$  tests on the 7-day retention session again showed that Groups 2, 3 and 4 had significantly fewer accidents than did the control group. More important, however, was the determination that those subjects who had experienced the simulated accident (Groups 3 and 4) had significantly fewer accidents than had Group 2 who were exposed to only a demonstration of the simulated accident, and that a transfer of training had

taken place. In order to locate the reliable differences between training and retention, further *t* tests for related measures were performed. The results of these tests indicated that the only significant changes over the retention period were increases in "accidents" in Group 1 ( $t = 2.592, 7 df, p < .05$ ) and Group 2 ( $t = 4.58, 7 df, p < .05$ ), the groups which were not subjected to the simulated accident.

### DISCUSSION

On the basis of this experiment, it was apparent that a single accident simulation had reduced the occurrence of potential accidents over a period of 7 days, even when tested in a different location.

### EXPERIMENT II

This study represented a replication and extension of Experiment I. The extensions were the inclusion of a retention test at 4 weeks as well as 1 week, increased group size, using female subjects as well as male and the use of a nonmodified pedestal grinder in the 4-week retention test.

### METHOD

#### Subjects

Seventy-two college students from an introductory psychology course served as subjects for this experi-

ment. Twenty males and 52 females between the ages of 19 and 23 years were randomly assigned to conditions, the only restrictions being that 5 males and 13 females comprise each group.

#### Apparatus

The apparatus for Experiment II was the same as that used for Experiment I except that an unmodified pedestal grinder was used in the second retention test.

#### Procedure

The group labels and the training procedures used for the groups were the same as in Experiment I. For the first retention test, each subject returned to the experiment exactly 1 week later and spark tested the 10 steel bars. Each subject also returned exactly 4 weeks after the original trial and again spark tested the 10 bars while the experimenter noted "accidents." After the training day no additional instructions, demonstrations, or experience with the simulated accident occurred.

### RESULTS

A preliminary analysis of the data with an *F* test revealed no significant difference between male and female subjects, so the data were combined on all further analyses. The analyses applied to the data of the second series were the same as those previously used.

Figure 2 shows the mean errors of each of the groups during the training and retention sessions. Again, it can be observed that the mean number of "accidents" decreased from Group 1 to Group 4 except that a reversal occurred between Group 1 and Group 2 at the second retention session.

The results of a two-way analysis of variance on the means of these errors showed that both the training procedures and the sessions variables were significant ( $F = 3.15, 4.59, 1.13; 3/68, 2/136, 6/136 df; p < .05, .05, ns$ ) but that their interaction was not.

The comparison of the mean number of accidents for each of the groups during each session was again made, using *t* tests. The results of the first session tests showed that all groups had fewer accidents than the control group (Group 1) and that Group 4, which was subjected to both the demonstration and experience of the simulated accident, made significantly fewer errors than Group 2, which had received only the demonstration of the simulated accident. The same results were ob-

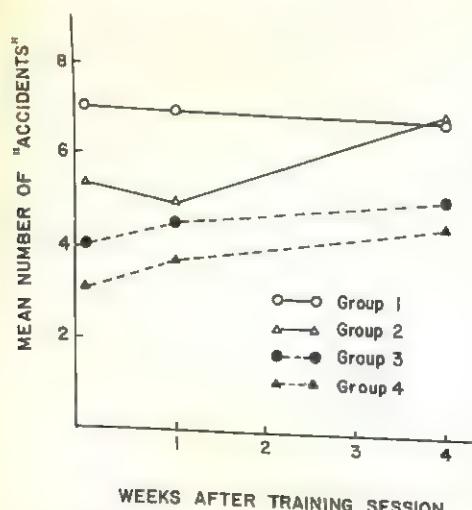


FIG. 2. Mean number of "accidents" for subjects within training conditions during training and two retention sessions ( $N = 72$ ).

tained in the analysis of the 1-week retention data, but the analysis of the data from the 4-week retention test revealed that the subjects in Group 4 had significantly fewer "accidents" than those in Groups 1 and 2, and although Group 3 had a lower mean number of accidents than had Groups 1 and 2, the difference was not statistically significant.

### DISCUSSION

The results of Experiment II closely parallel those of Experiment I in most respects. Again a single experience or demonstration of a simulated accident during training significantly reduced the number of "accidents" during the training and the first retention session 7 days later.

However, the data of the second retention session, 4 weeks after the training session, show that the effect of the training procedures was diminishing—the only group with significantly fewer "accidents" than Group 1 (control) being Group 4. This finding is, of course, more to be expected than was the trend toward continued reduction of errors in the retention session found in the first experiment.

That the experience of the simulated accident is more effective than its demonstration can be inferred from the order of the mean number of "accidents" and the data that show the number of accidents in Group 2 (demonstration of simulated accidents only) increased significantly over the retention period in both experiments while the "accidents" in Group 3 (experience of simulated accidents only) remained low.

It is interesting to speculate as to why Group 4 was superior to the other groups in both experiments and was the only group to retain its advantage after 28 days. Group 4 was the only group to receive both a demonstration and the experience of a simulated accident. In other words, the subjects of Group 4 had a double exposure to the accident simulation. Thus, the number of exposures to the simulated accident may be the most important parameter in the effective use of accident simulation in teaching safe operating procedures.

TABLE 1

NUMBER OF SIMULATED ACCIDENTS, CONSECUTIVELY OR INTERMITTENTLY PRESENTED ON PRESELECTED TRIALS FOR EACH TREATMENT GROUP

Group	Number of simulated accidents	Schedule	Trials on which accidents occurred
A	2	Consecutive	1, 2
B	2	Intermittent	5, 11*
C	5	Consecutive	1-5
D	5	Intermittent	4, 6, 7, 9, 10*
E	10	Consecutive	1-10
F	10	Intermittent	1, 4-8, 11-14*
G	0	(Control)	
H	0	(Control)	

\* Randomly selected.

### EXPERIMENT III

Experiment III investigates the effect of varying the number and pattern of presentation of the simulated accidents during the training session on the number of "accidents" occurring in three retention sessions up to 6 months later.

A  $4 \times 2 \times 3$  factorial design was employed with 0, 2, 5, and 10 simulated accidents, using two different schedules (consecutive and intermittent) of accident presentation and retention trials at intervals of 1, 3, and 6 months.

It was hypothesized that an increase in the number of simulated accidents during training would result in fewer accidents during retention. Also, the considerable research on intermittent reinforcement and its resistance to extinction (Jenkins & Stanley, 1950) led to the hypothesis that the intermittent schedule of accident presentation would be more effective than the consecutive in maintaining safe procedures.

### METHOD

#### Subjects

One hundred and twenty male and female college sophomores volunteered to be subjects. The age of the subjects ranged from 18 to 22 years. The subjects were randomly assigned to groups, with the limitation that the proportion of males to females were to be the same in each group. Nine females and six males were in each of the eight groups at the start of the first phase.

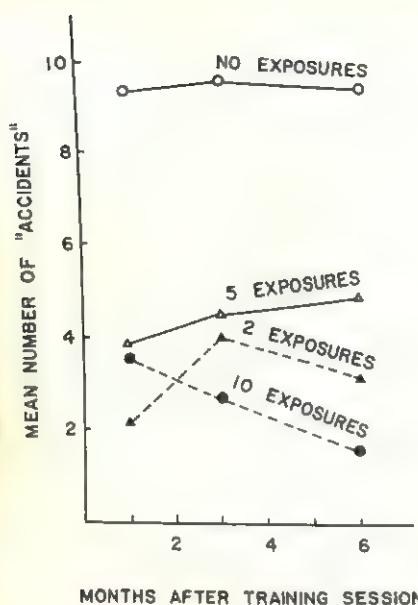


FIG. 3. Mean number of "accidents" for subjects within training conditions as a function of three retention sessions ( $N = 80$ ).

#### Apparatus

The apparatus was the same as that used in Experiments I and II. The training session and all the retention sessions were performed on the same grinder in the same location. The subjects received 15 trials on the cover task each session. The groups and their treatment during the training session are presented in Table 1. All groups were treated identically during the retention sessions.

#### RESULTS

Figure 3 shows the mean number of "accidents" that occurred in each group over the retention intervals. The  $F_{max}$  statistic (Winer, 1962) was used to test for homogeneity of variance. There were two cells sufficiently deviant to yield a significant ratio ( $F = 12.15$ ,  $9/24$  df,  $p < .05$ ). In that heterogeneity must be quite extreme to be of serious consequence (Norton, 1953), an analysis of variance was performed.

A three-way analysis of variance, mixed model, was performed on the "accident" data. The three main variables were number of accidents, distributed versus continuous accident exposure, and retention intervals. Since a substantial number of subjects resigned during the 6 months of the experiment, it should be noted that in order to equalize the  $N$  of each

group, subjects were randomly excluded from all groups until their  $N$  equaled the  $N$  of the smallest group, which by the end of the last phase was 10. Thus, the data for 10 subjects for each treatment group, and a total  $N$  of 80 was used for each of the three retention trials.

The results of the analysis revealed that only the number of simulated accidents variable was significant ( $F = 17.33$ ,  $3/126$  df,  $p < .01$ ). The subjects were, therefore, grouped according to the number of simulated accidents, and subgroups based on the accident schedule were ignored.

In order to probe the nature of the differences between the treatment totals following the significant overall  $F$ , the Newman-Keuls procedure (Winer, 1962) was used. The results of this analysis indicate that the control group (no simulated accidents) was significantly different from each of the simulated accident groups, but there was no significant difference between any of the simulated accident groups.

#### DISCUSSION

The results of Experiment III strongly support those of the two preceding experiments in showing that under the conditions used, the results of one training session involving accident simulation is effective in promoting safe tool operation 6 months later.

Contrary to expectations, the retention tests revealed little difference in number of accidents between those trained under the consecutive accident-simulation condition and those trained under the intermittent condition or groups given 2, 5, or 10 simulated accidents. It is possible that training under intermittent conditions may produce longer lasting effects than the consecutive condition, but the 6-month period allotted for forgetting in this study was not sufficient to show the difference. Over a longer period, one might then observe differences in retention curves between the groups.

The results of these studies strongly indicate that the use of accident simulation as a training method for the safe operation of a power tool is a powerful technique.

It is suggested then, that the following procedures may be the basis of an effective

training program for the safe operation of power tools.

1. Any unsafe acts associated with the power tool be identified.

2. A simulated accident that could result from the unsafe acts be devised and suitable equipment to simulate them be installed on the power tool.

3. The trainee should then be allowed to operate the equipment and be subjected to the simulated accident when he performs the unsafe act.

4. If a trainee does not perform an unsafe act during the training session, the simulated accident should be demonstrated.

Since all the subjects in the present experiments were inexperienced in the use of the tool under investigation and presented no previously developed habit patterns in the use of the tool, an important extension of this work should be the determination of what effect, if any, accident simulation would have in altering already established unsafe habit patterns.

#### REFERENCES

GIBSON, J. J. The contribution of experimental psychology to the formulation of the problem of

safety—A brief for basic research. In, *Behavioral approaches to accident research*. New York: Association for the Aid of Crippled Children, 1961.

HADDON, W., SUCHMAN, E., & KLEIN, D. *Accident research*. New York: Harper & Row, 1964.

HEINRICH, H. W. *Industrial accident prevention*. New York: McGraw-Hill, 1950.

JENKINS, W. D., & STANLEY, J. C. Partial reinforcement: A review and critique. *Psychological Bulletin*, 1950, 47, 193-234.

NORTON, D. W. An empirical investigation of some effects of nonnormality and heterogeneity on the F-distribution. Unpublished doctoral dissertation, University of Iowa, 1952. Cited by Lindquist, E. F., *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

RUBINSKY, S., & SMITH, N. Evaluation of accident simulation as a technique for teaching safety procedures on small power tools. (Research Rept. No. RR-70-4, Contract No. PH 86-68-607) Providence, R.I.: Injury Control Research Laboratory, 1970.

UNITED STATES DEPARTMENT OF LABOR, BUREAU OF LABOR STANDARDS. *Safety Standards*, 1971, 20(2), 14.

VAUGHN, J. *Positive versus negative instruction: An experimental study of the effects of various types of instruction on behavior*. New York: National Bureau of Casualty & Surety, Underwriters Educational Service, 1928.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received August 24, 1971)

## INFORMATION SEEKING WITH CONFLICTING AND IRRELEVANT INPUTS

JERROLD M. LEVINE<sup>1</sup>

*American Institutes for Research, Washington, D.C.*

Information-seeking performance was studied under conditions of conflicting and irrelevant input information in an eight-choice task that was an abstracted version of a tactical decision problem faced by military commanders. Fourteen college students were required to purchase information from three fallible sources until they could decide which target was the object of an enemy advance. The earlier a correct choice was made, the greater the monetary payoff to the subject. The results indicated that degree of information conflict and relevance had little influence on trial number and latency of correct choices, but a more marked impact on initial decisions. Subjects purchased more information prior to first decisions when degree of relevance was low. Choice latencies of first decisions decreased with increasing relevance and decreasing conflict.

The development of advanced military command and control systems and the increasingly important role of intelligence information in tactical decision making has emphasized the role of man as a sequential processor of data. A common task involves the gathering of information for an ultimate choice among alternative courses of action. At each point in time, a decision must be made whether to choose a terminal course of action based on current information or whether to gather additional information. Postponing ultimate action may further reduce the uncertainty associated with the choice of actions but it may risk delaying action so long as to inhibit its effectiveness. The behavior generated by such tasks is known as information seeking or optional stopping. The critical decision is when to stop acquiring information.

When a sequential task such as described is considered in the context of intelligence systems containing multiple information sources, it is possible that information will at times be conflicting. The intelligence analyst, for example, may receive photointerpreter reports of the presence of an enemy tank company in a geographical area, while ground patrols report no such activity, and prisoner interrogations, on the other hand, lead to a different conclusion. The impact of these pieces

of information will, of course, depend on the credibility and variability of the source as well as other factors. One purpose of the present study was to evaluate the effect of degree of conflict among information sources on information-seeking performance.

The decision to seek additional information prior to making a terminal decision always involves a cost in terms of either or both of the tangible and intangible resources of the decision maker. Such costs encourage the decision maker to make a terminal decision as soon as possible. On the other hand, the more resources expended, the more information there will be on which to base a decision. The information obtained from any particular source, however, may not always aid the decision maker in selecting his choice of action. There is not usually a guarantee that information will be relevant to the decision that must be made. This is especially true when the time of arrival and nature of the information is not under the control of the decision maker. Indeed, no information at all may be available from a source when it is queried. The second purpose of this study was to evaluate the influence of degree of information relevance on information-seeking behavior.

Although a substantial amount of research on information seeking has been conducted, the influence on performance of conflicting and irrelevant inputs has not been studied.

<sup>1</sup> Requests for reprints should be sent to American Institutes for Research, 8555 Sixteenth Street, Silver Spring, Maryland 20910.

Much of the past work has been concerned with the statistical parameters of the input information (e.g., Becker, 1958; Howell, 1966; Irwin & Smith, 1956), payoff (e.g., Edwards & Slovic, 1965; Rapoport & Tversky, 1966; Irwin & Smith, 1957; Pitz & Reinhold, 1968), and an evaluation of strategies of information seeking in terms of optimal models (e.g., Fried & Peterson, 1969; Pitz, 1968, 1969). The present study differs from most previous research in several respects. First, the task was eight-choice and, therefore, more difficult than the two- or four-choice problems used by most investigators. Second, the diagnostic value of each sequential piece of information was not equal in the present study as is the case in most past efforts. In the present experiment, data samples provided in late stages of the task contain more information than early data since late reports of enemy activity are more predictive of the enemy objective than are early reports. Third, latencies for decisions to sample information as well as for decisions to select an alternative were measured.

### METHOD

The experimental task was an abstracted version of a tactical decision problem faced by a battlefield commander similar to that of Hammer and Ringel (1965). The task required that subjects determine which one of eight friendly locations was the target of an enemy advance. The enemy advance took place in 48 discrete steps and was displayed to subjects as a "pathway" initiating at the bottom center of a screen. The display depicted the current enemy position as well as the entire past history of the advance. Each subject requested updated position reports as he deemed necessary until he was ready to decide which target was under attack. The earlier he made a correct choice, the more money he earned.

### Apparatus

The stimulus materials were rear projected onto a 24-inch square screen by a Kodak Carousel slide projector. The projector was connected to two sets of control switches and a master console. The master console consisted of digital logic circuitry, interval timers, a Beckman EPUT counter, and a Friden motorized tape punch. The experimenter's control box served simply to activate or deactivate the system. The subject's control box had three button-switches and associated indicator lamps. The buttons were labeled Information (I), Decision (D), and Restart (R). The I and R buttons when depressed stopped the EPUT counter, allowed for response

latencies to be punched out, reset the counter, advanced the slide projector, and started the EPUT counter again. The D button performed the same functions except for the slide advance. An event counter placed on the screen kept the subjects aware of the number of updates received. The entire system was duplicated so as to permit two subjects to be run simultaneously. Gaussian noise was presented through earphones to the subjects in order to prevent them from hearing the projectors operate.

### Subjects

Fourteen male undergraduate students from local universities served as experimental subjects. All subjects were paid for their services.

### Stimulus Materials

The stimulus materials consisted of sets of 35 mm. color slides. Each slide depicted the location of the hypothetical enemy advance and the entire past history of locations of the enemy as reported by each of three information sources. The geographical points were connected by colored straight lines, each color representing a different source. The first slide of a set showed the enemy reported to be at the bottom of the screen by all three information sources. As additional information was provided, each source generated a pathway toward one of the eight targets at the top of the screen. The last slide of a set showed the three pathways converging on one of the targets. Forty-eight slides comprised a single set defining one problem. Twenty-four sets of slides were constructed in the following systematic manner: A  $48 \times 48$  unit square matrix was prepared with the eight targets equally spaced along the top of the matrix and trials (steps of the enemy advance) depicted along the ordinate. There were 48 such trials, and these were divided into eight blocks of six steps each. The aggressor pathway suggested by one information source was constructed by drawing a straight line from the bottom center of the matrix to a preselected target. An isosceles triangle with a  $20^\circ$  apex at the target was constructed about this line. Straight lines at angles either less than, equal to, or greater than  $90^\circ$  to the abscissa of the matrix were then drawn within each block of six steps. These lines were connected to form a random fluctuation about the altitude of the triangle. The length of the within-block lines was limited by the sides of the triangle which also limited the variability of the pathway such that it constantly decreased and converged on the target.

Three such triangles were generated. The center one was always as described above. The other two triangles and their respective pathways were situated to the left and right of the center one and overlapped it. The apices of each triangle were located at the same point, thus giving each pathway a different starting point. Within corresponding blocks for the three triangles, the direction of the lines either all agreed or none agreed or two of the three agreed. This agreement in direction of within-block lines

defined the experimental conditions of conflict. Step-by-step enemy locations were distributed randomly about these lines with the restriction that the line was the best fitting one around the points. The points were connected by colored tape to establish the final pathway to be photographed. Conditions of information relevance were established by randomly removing some of the lines within a block.

### *Experimental Design*

The independent variables were degree of conflict among information sources, degree of information relevance, and variability of the pathways. Degree of conflict was defined operationally in terms of the direction of the least-squares lines about each of the three information sources data points within blocks of six steps. If all three information sources depicted the same direction, there was 0% conflict. If all three disagreed, there was 100% conflict, and if one source disagreed with the other two, there was 33% conflict. Degree of relevance of information was defined in terms of the proportion of instances that all three sources provided information. When updated information was requested, all three sources either provided information or they did not. Within a block of six steps, the sources provided information either on three, four, five, or six occasions thus defining 50%, 67%, 83%, or 100% conditions of information relevance. Variability of the pathways was defined in terms of the variance of each of the three pathways about a straight line to the target. When one third of the data points of each pathway overlapped those of an adjacent pathway, there was high variability. When there was two thirds overlap, low variability was said to be present. The overlap was manipulated by changing the separation of the three triangles and was independent of degree of conflict.

A  $3 \times 4 \times 2$  factorial design was generated. All subjects were required to perform in each combination of conditions. Each of the 24 conditions had a unique problem associated with it. Targets were assigned to conditions so that the conflict conditions were orthogonal to targets, but relevance and variability conditions were partially confounded with target position.

### *Procedure*

Subjects were briefed on the general nature of the task and instructed that their objective was to decide which of the eight locations was the target of an enemy advance. Information concerning the location of the enemy force was provided on request at a fixed cost (one unit of resource). Such a request either provided updated data from all three information sources or no data at all. As each piece of information was provided, the subject had to decide whether to purchase additional information or to stop and decide which target was the object of the enemy advance. Subjects were advised that the position reports from the three information sources might conflict in terms of location and direction of

movement since the enemy was taking evasive actions, and the information reports were fallible. Correct decisions were rewarded and incorrect ones were penalized in a manner consistent with the military situation being simulated; i.e., the longer it took to reach a decision, the less effective the action based on that decision would be and, therefore, the smaller the payoff to the subject.

Subjects requested updated information by depressing their information button in a self-paced fashion. Each such action advanced the enemy pathway and permitted the recording of the subject's information sampling latency. When the subject thought he knew which target was under attack he depressed his decision button and recorded his decision on prepared forms. Subjects expressed their decision by entering probabilistic estimates of each of the eight targets being the actual one under attack. The target assigned the highest probability represented their choice of the correct target. Decision latencies for this response exclusive of writing time were recorded automatically. The subject then depressed his restart button in order to advance the slide projector and receive an updated report. Feedback was provided by requiring subjects to go through all 48 slides and thus determine which target the pathways ultimately converged on. If the subject was correct at the trial upon which he made a choice, he was not charged for the additional information sampled. If, on the other hand, as he sampled additional information, the subject came to believe that his initial decision was incorrect, he could change his choice by re-depressing the decision button and indicating his new choice. In this case, he was charged for the additional information. Subjects were permitted to make up to three decisions.

Subjects performed in pairs for approximately 2 hours in each of four sessions spaced across a week. The first session was devoted to briefing and training problems. The next three sessions each required subjects to work on an average of eight test problems. The 14 subjects were run within a 2-week period. Each test problem required an average of 5 minutes to complete. Problems were assigned to subjects in random order.

### *Payoff*

The payoff schedule has the following characteristics:

1. The earlier a correct decision, the greater the payoff.
2. The earlier an incorrect decision was corrected, the greater the payoff.
3. Each incorrect decision made after a correct one reduced total payoff by 20%.
4. Incorrects made prior to a correct were penalized only in terms of the number of additional information updates required before a correct choice was made.
5. The trial number of a correct response weighted the payoff function in a nonlinear fashion so that

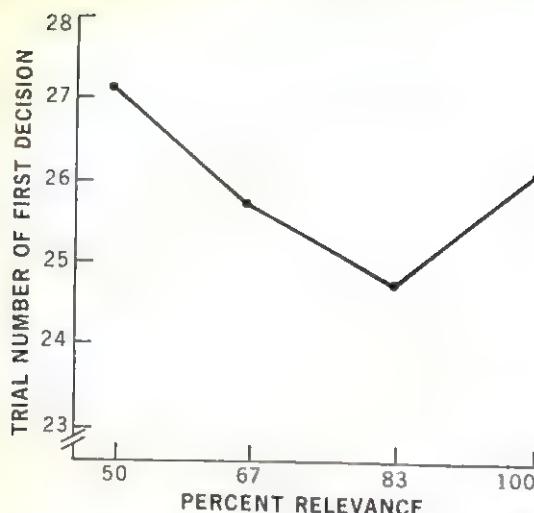


FIG. 1. Trial number of first decision as a function of degree of information relevance.

late corrects were rewarded considerably less than early corrects.

6. Subjects could not experience a loss.

The task and payoff function were designed to permit subjects to modify their initial decisions if additional information led them to change their minds. They could earn from 1 to 95 units of resources on each problem. Each unit was worth \$.02. The payoff function was explained in detail and examples provided. During the six training problems, payoff was computed immediately following each problem and provided to the subjects. During the 24 test problems, subjects received no feedback concerning their earnings.

## RESULTS AND DISCUSSION

Performance was evaluated with regard to information-seeking, latency, and confidence scores for both initial decisions and correct decisions. Unless otherwise noted, all results were based on analyses of variance.

Relevance of information proved to reliably influence the trial number of a first decision [ $F(3,39) = 4.25, p < .025$ ]. Initial decisions were reached earlier in the sequence of events as the proportion of relevant information increased. This relationship, however, was not maintained for the 100% relevant information condition as shown in Figure 1. There were no significant differences among conditions of conflict or variability of information, nor were any interactions statistically reliable.

An analysis of choice latencies for initial decision indicated a significant conflict by

variability interaction [ $F(2,26) = 3.77, p < .05$ ]. Under low-variability conditions, decision times were greatest for partially conflicting data, while for high-variability conditions partially conflicting information resulted in lower decision times than the two other conflict conditions (see Figure 2). No other reliable effects were obtained. The data did suggest, however, that latencies decreased with an increase in degree of information relevance and a reduction of information conflict.

An evaluation of the amount of information purchased prior to a correct decision revealed that only the conflict by relevance interaction was statistically reliable [ $F(6,78) = 2.83, p < .025$ ]. A plot of the means comprising this interaction, however, failed to show any clear functional relationship. Averaged across all subjects and experimental conditions, the mean number of inputs purchased prior to a correct decision was 31.77 as compared to an average of 25.96 information requests prior to a first decision.

An analysis of decision latencies for correct choices indicated no significant differences as a function of levels of the independent variables. The mean decision time averaged over subjects and conditions was 8.35 seconds. This was 13% faster than the 9.51 seconds required to make an initial decision.

Analyses of variance were carried out on probabilities assigned to the target choice and to the correct target for initial decisions and correct decisions. No significant sources of

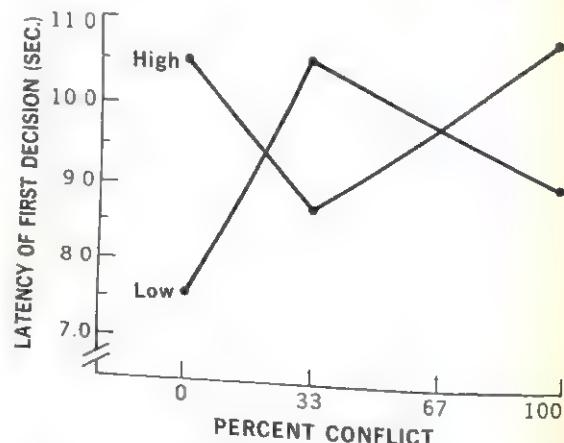


FIG. 2. Latency of first decision as a function of degree of information conflict with variability as the parameter.

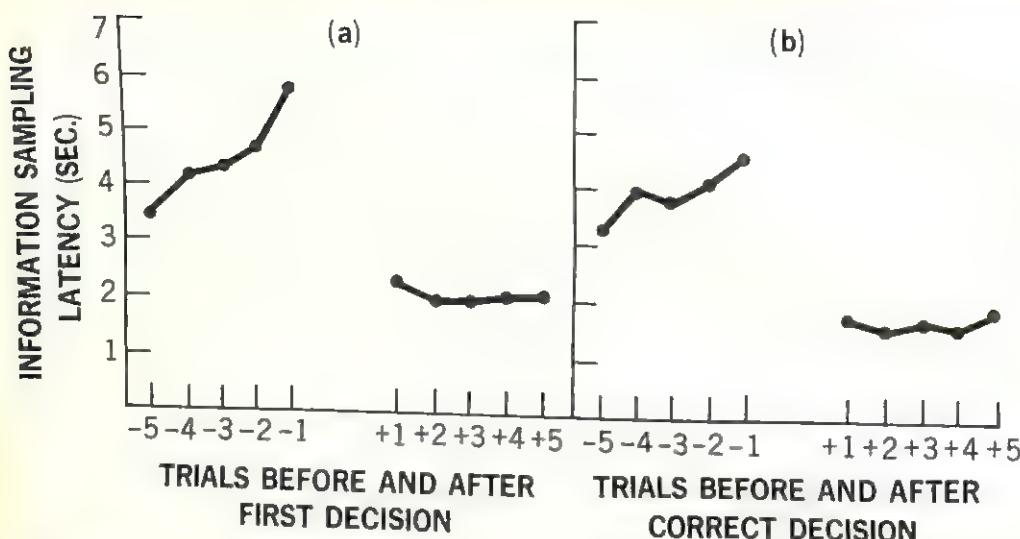


FIG. 3. Information sampling latency on trials before (-) and after (+) the first decision (a) and the correct decision (b).

variance were detected. When the subjects were correct, their average confidence was only 61%, indicating that they were still unsure of the target under attack. The mean subjective confidence in the choice made on initial decisions was 54%, suggesting that subjects sampled information prior to a first decision until the subjective probability of one target being correct was just greater than the cumulative subjective probability of all other targets being correct. The subjects did not await any major increment in their certainty between first and second decisions. The expressed confidence in the correctness of their choice averaged only 9% higher for second decisions than for first decisions. Apparently subjects were determined to make early decisions of which they were relatively uncertain rather than to experience the costs associated with delayed decisions.

Mean information sampling latencies for each of the five trials prior and subsequent to initial and correct choices were computed. These distributions, when plotted separately for the levels of each independent variable, indicated no discernible differences as a function of experimental conditions. The distributions were therefore averaged over conditions and are presented in Figure 3. Average information sampling latency increased linearly as the decision maker approached a

choice point. Subsequent to a choice, sampling latencies were markedly reduced and were nearly constant across further trials. These relationships were evident when considering the trial number of a first choice or the trial number of a correct choice. Apparently subjects considered information inputs more carefully with each successive input just prior to making a choice.

Mean values for sampling information based on the five trials before and after the trial of a choice were computed and indicated that decisions to stop sampling information and make a choice among alternatives required more than twice the time needed for decisions to request additional information prior to a choice. These data suggest that subjects made more than a simple decision to stop sampling information since such a decision would not be expected to require more time than the decision to seek information. It may be that subjects decided to stop and, in addition, choose an alternative, a two-fold decision, before responding by depressing their decision button. On the other hand, since decisions to stop sampling information in effect commit the subject to a choice which might be costly if wrong, while decisions to continue sampling involve a relatively small cost, these latency differences may indeed reflect the fact that more time is required for riskier deci-

sions. No conclusions can be drawn from the available data.

Each piece of information purchased subsequent to an initial decision was potentially more costly than information purchased prior to that decision, if the decision proved to be incorrect. Since subjects, on the average, were only 54% confident that their choice was correct, it would be to their advantage to sample inputs more cautiously after initial decisions. That this was not the case is evident from Figure 3. Sampling latencies were considerably shorter after initial decisions than before. This behavior, while not consistent with expectations based on the payoff function, is quite reasonable when one considers that several pieces of information subsequent to a decision were necessary before that decision could be verified. Apparently, subjects learned that one or two inputs beyond the point at which they initially made a decision did not provide enough additional information to attempt a reevaluation of that decision.

In order to assess the possible influence of target position and the interaction of position with degree of conflict, analyses of variance were carried out on the amount of information purchased prior to both initial and correct responses. Since target position was partially confounded with relevance and variability, no assessment of these interactions was possible.

The results indicated that target position was significant ( $F(7, 91) = 2.28, p < .05$ ) for the amount of information purchased prior to an initial decision, but not for the amount purchased prior to a correct decision. The significant position effect was further evaluated by testing the hypothesis that decisions would be reached with less information when pathways converged toward targets positioned at the left or right extremes of the display, than when targets were positioned in the center of the display. Such a hypothesis would suggest that problems for which the true target was at the extreme were easier to solve than problems for which the true target was in the center, since there were fewer possible target alternatives to be considered at the extremes than at the center of the display. Tukey tests, however, did not support this

hypothesis. Thus, the performance difference found as a function of target position did not suggest that the extremes and center positions differed.

The interaction of target position and degree of conflict proved to be statistically reliable ( $F(14, 182) = 2.10, p < .05$ ) only for the amount of information purchased prior to a correct decision. An evaluation of the extreme versus center positions suggested that more information was purchased prior to a correct decision for targets positioned at the extremes of the display than for center targets, when there was complete conflict, but this relationship was reversed under partial conflict conditions. For the no-conflict condition, performance was essentially equal regardless of target position. This interaction and the possibility of other target position by independent variable interactions which could not be tested for, raise the possibility that the effects of the independent variables may have been partly accounted for by target position. However, due to counterbalancing of target positions with the independent variables, any such interaction is likely to be marginal.

#### CONCLUSIONS

While there is some suggestion that conflicting and irrelevant inputs as manipulated in this study influence information-seeking performance, no clear-cut relationships proved to be statistically reliable. In fact, considering trial number and latency of correct choices as performance measures, the suggestion was that subjects quite capably integrate information regardless of its degree of conflict, relevance, or variability. While it appears that subjects seek more information prior to initial decisions when degree of relevance is low, correct decisions are reached with the same number of information requests under all conditions of relevance. Likewise, latencies of first decisions appear to decrease as degree of relevance increases and degree of conflict decreases, but latencies of correct choices are not influenced by these variables. It appears that the influence of the independent variables diminished as the subject progressed through each problem.

The analyses of subjects' confidence ratings were quite surprising. Intuitively, one would expect subjects to be more confident in the correctness of their decisions under conditions of decreased variability and degree of conflict and increased relevance of input information. That this was not the case may imply that independent of the conditions under which subjects were required to perform—they developed some preconceived notion about the quantity of information that should be sampled. If this speculation is correct, subjects may have based their assessments of confidence on the amount of data available rather than on the predictability of the correct target given that amount of data.

The absence of previous research on the effects of information conflict and relevance on information seeking suggests that the present results be viewed as preliminary findings. It is particularly important to study these parameters of information seeking with additional displays and under modified operational definitions of the variables in order to establish the generalizability of the present results. The manner in which conflict and relevance were defined and manipulated in the present study may not have been maximally perceived by subjects. If these variables could be made more manifest, their impact on performance might be enhanced. This is particularly true of the relevance variable. Pathways containing missing data were depicted graphically, and subjects might have easily interpolated between missing points, thereby mitigating the influence of this variable upon information-seeking performance. Additional aspects of the present study which must be investigated in future research include: (a) the costs of data sampling relative to payoff (in the present study this ratio was low, resulting in an indiscriminate sampling of inputs), (b) the risk factor imposed by the payoff function (in the present study sub-

jects could not experience a monetary loss on any problem), and (c) the influence of differential diagnosticity of purchased information.

The accomplishment of these suggested research efforts and others would provide sufficient information on the effect of conflicting and irrelevant inputs on information seeking so that information processing systems could be developed and training in information integration strategies provided in an attempt to maximize decision-making performance.

#### REFERENCES

- BECKER, G. M. Sequential decision making: Wald's model and estimates of parameters. *Journal of Experimental Psychology*, 1958, 55, 628-636.
- EDWARDS, W., & SLOVIC, P. Seeking information to reduce the risk of decisions. *American Journal of Psychology*, 1965, 78, 188-197.
- FRIED, L. S., & PETERSON, C. R. Information seeking: Optional versus fixed stopping. *Journal of Experimental Psychology*, 1969, 80, 525-529.
- HAMMER, C. H., & RINGEL, S. The effects of amount of information provided and feedback of results in decision making efficiency. *Human Factors*, 1965, 7, 513-519.
- HOWELL, W. C. Task characteristics in sequential decision behavior. *Journal of Experimental Psychology*, 1966, 71, 124-131.
- IRWIN, F. W., SMITH, W. A. S., & MAYFIELD, J. F. Tests of two theories of decision in an expanded judgment situation. *Journal of Experimental Psychology*, 1956, 51, 261-268.
- IRWIN, F. W., & SMITH, W. A. S. Value, cost, and information as determinants of decision. *Journal of Experimental Psychology*, 1957, 54, 229-232.
- PITZ, G. F. Information seeking when available information is limited. *Journal of Experimental Psychology*, 1968, 76, 34-75.
- PITZ, G. F., & REINHOLD, H. Payoff effects in sequential decision making. *Journal of Experimental Psychology*, 1968, 77, 249-259.
- PITZ, G. F. Use of response times to evaluate information seeking. *Journal of Experimental Psychology*, 1969, 80, 553-557.
- RAPORT, A., & TVERSKY, A. Cost and accessibility of offers as determinants of optional stopping. *Psychonomic Science*, 1966, 4, 145-146.

(Received July 28, 1971)

## AN EXPERIMENTAL INVESTIGATION OF THREE METHODS OF PROVIDING WEIGHT AND PRICE INFORMATION TO CONSUMERS<sup>1</sup>

ROBERT D. GATEWOOD<sup>2</sup>

*Department of Management,  
University of Georgia*

ROBERT PERLOFF

*University of Pittsburgh*

A supermarket setting was simulated to experimentally evaluate three methods of presenting information to consumers: (a) current supermarket method (showing total price and net weight); (b) current supermarket method, but adding a computational device to aid in price calculations; and (c) current supermarket method, but providing also price per ounce of net weight of product. Seventy-five volunteer subjects were used in a design where 25 subjects were assigned randomly to each of the three methods of presentation. The experimental task was for subjects to choose the most economical package for each of nine product groups. Results indicated that presenting the additional information of price per ounce of net weight produced a significant increase in accuracy of choices, while significantly reducing the time required to make such choices.

Truth in the packaging and the pricing of products in the American marketplace has been a subject of public controversy in recent years, despite the 1966 "Fair Packaging and Labeling Act." The basic issue in this controversy is alleged consumer confusion in the determination of price comparisons. The 1966 law was designed to reduce this confusion by (a) regulating the location, the print size, and the statement of net contents; (b) directing the establishment of standard definition of such terms as "serving" and "small," "medium" and "large" sizes; and (c) empowering certain agencies, in extreme conditions, to establish the net weights of packages and number of sizes to be offered for a product group.

Recently, several consumer advocates have criticized the effectiveness of this law on the grounds that price comparisons are no easier for the consumers to accurately make now than before enactment. The Consumer Federation of America (Cohan, 1969) has said, "Truth in packaging . . . is one of the best non-laws on the books [p. 10]." Also, Virginia

Knauer (Mrs. Knauer Twits Commerce . . . 1969), Presidential Advisor on Consumer Affairs, has stated, "We don't think the labeling on products has adequate or clear information. We think something should be done about the Fair Packaging and Labeling Act [p. 1]."

The most frequently proposed alternatives are (a) to equip consumers with small devices which, when given total price and weight, would yield price per unit and (b) to require retailers to clearly mark the price per unit on each item. Grocers argue that devices such as in *a* above are extremely easy to use, have universal application, and require neither additional change in the law nor great expense to implement. The second alternative (in *b* above) is favored by many consumer advocates as being more effective in providing necessary information to consumers for making price comparisons. Grocers, in general, have opposed this alternative. *Advertising Age* (Grocers Moan . . . 1969) has written, ". . . supermarket managers and suppliers complain that such a regulation will cause them to double their labor force, raise prices, or go out of business altogether [p. 3]."

In the present article, an attempt is made to evaluate the consumer's ability to process weight and price information in making

<sup>1</sup> This study is based on a doctoral dissertation submitted to Purdue University in partial fulfillment of the requirements for the doctoral degree.

<sup>2</sup> Requests for reprints should be sent to Robert D. Gatewood, Department of Management, College of Business Administration, University of Georgia, Athens, Georgia 30601.

price comparisons under the foregoing two methods as well as under the present supermarket method. It would seem critical to collect such information prior to implementation of one of these methods nationally.

Previous research in this area is extremely limited. A survey of the psychological literature indicates only one experimental investigation prior to the passage of the 1966 act. Friedman (1966) directed 33 young home-makers, each having completed at least 1 year of college, to select the most economical (largest quantity for the price) package for each of 20 supermarket products. Of the 660 purchased, 284 (43%) were purchased for more than the lowest price, indicating that these subjects could not adequately process this information.

The research reported in the present article tested the following hypotheses:

1. Consumers are significantly more accurate in choosing the most economical package from a product group when price-per-unit information is directly available for this product group than when the information is not available.
2. The time required for consumers to choose the most economical package is significantly less when price-per-unit information is directly available than when it is not directly available.
3. There is no significant difference in accuracy of consumer choice of most economical packages between the following two experimental conditions: (a) current display methods and (b) current display methods with the addition of a computational aid.
4. There is no significant difference in time required to arrive at decisions of most economical package between the two conditions stated in Hypothesis 3.
5. The number of sizes of packages within a product group is significantly negatively correlated with the number of correct choices of most economical package for the two experimental conditions: (a) current display methods and (b) current display methods with the addition of a computational aid. For the condition of price-per-unit information, the correlation will not be significant.

## METHOD

A simulated supermarket situation was set up to collect data for this research, using samples of nine food products as experimental items. The simulation was meant to be representative of the shopping situation that a consumer is confronted with in a supermarket. Therefore, the nine product groups and the items within each were samples drawn from a single supermarket. This supermarket was a member of a large chain that was judged to stock approximately the same products, sizes, and brands as other members of the chain.

The nine product groups chosen for experimentation were randomly selected from the nonperishable items carried by the supermarket. Specifically, the sampling was random selection from within each size within a product group.

Seventy-five volunteer subjects participated in this investigation, 64 of whom were women; 60 of the 75 subjects had completed at least 1 year of college; 48 were between 20 and 29 years of age, 17 were between 30 and 39, and 10 were 40 years or older.

Subjects were assigned randomly to the three treatment conditions, 25 to each. Each subject performed the experimental task individually. When volunteers reported for the experiment, all were told that their task was to choose the "most economical package" for each of the nine product groups; this was defined as using the information available on the food packages (weight, servings, strength, etc.) to choose that package which gave the most quantity for the money, or the "best buy." All food packages used in the simulation were numbered. To indicate his choices of most economical packages, a subject was asked to write only the numbers of the packages of his choice on the answer sheet he was provided.

In Treatment A, the subjects were presented with the same information as in a supermarket. That is, the packages were presented with the net weight and/or the number of servings on the label and the total price stamped on the package. Quantity and price were displayed in the same manner for Treatment B. However, the subjects in this condition were asked to make use of a computational aid to assist them in making decisions. The device requires the consumer to match the total price of the package (recorded on the outer circle of the device) with the net weight of the package (recorded on the movable inside wheel of the device); the cost per ounce of the package is then shown in a box in the center of the wheel. All subjects in Treatment B were instructed in the use of this device and trained to a criterion of three successful price-per-unit computations. For Treatment C, subjects were presented with quantity and price as in Treatment A and, in addition, with cost per ounce of net weight of the package. This method of presenting information is commonly referred to as "unit pricing." This information was calculated and printed on small slips of paper which were placed under each package.

A major question in any simulation experiment concerns the fidelity of the simulation and whether the subject behavior elicited under the simulated conditions is representative of the behavior under actual conditions. To estimate the representativeness of the experimental behavior of the subjects, 12 additional subjects were asked to complete the experimental task in the supermarket from which the food items were purchased. It was possible to have subjects "shop" under Conditions A and B in an actual supermarket because these conditions did not require any change in the normal method of presentation of information by the supermarket. Twelve additional subjects therefore were asked to complete the same experimental task, six for each of these two treatments, as were the subjects in the simulation constituting this experiment.

### RESULTS

Single-factor analyses of variance were performed comparing the three treatment conditions on accuracy of choice of most economical and total time required to make all nine choices.

The scoring key was determined for all product groups, except one, by dividing the net weight of each package into the total price of the package, yielding cost per ounce. For these eight product groups, the correct answer was the package having the lowest price per ounce. For one product group, instant potatoes, the most economical package was determined by dividing the number of ounces of potatoes made by the contents into the total price of the package; it was this cost per ounce that was the correct choice, and not the cost per net weight. Information regarding the number of servings and size of servings (defined on all packages as 4 ounces) was conspicuously displayed in all packages and assumed to be accurate. It should be noted that subjects were instructed that the most economical package should be determined from information presented on the package and could be in terms of net weight, servings, or in any other conventional measure.

The mean number of correct choices for each of the three treatments was A, 5.72 ( $\sigma = 1.31$ ); B, 5.96 ( $\sigma = 1.57$ ); and C, 8.04 ( $\sigma = .45$ ). An ANOV yielded an overall  $F$  value of 27.00, significant beyond the .01 level ( $df = 2/72$ ). The Newman-Keuls test for probing the nature of the differences between treatment means following a significant over-

all  $F$  indicated the differences between Treatment Groups A and C and B and C to be significant ( $p < .01$ ). The difference between Treatment Groups A and B was not significant.

The mean number of minutes spent in making the nine choices for each treatment was A, 23.93 ( $\sigma = 10.00$ ); B, 31.72 ( $\sigma = 9.57$ ); and C, 3.60 ( $\sigma = 1.11$ ). An ANOV yielded an overall  $F$  of 111.4, also significant beyond the .01 level ( $df = 2/72$ ). Newman-Keuls analyses indicated all differences among the three treatment groups to be significant ( $p < .01$ ) with the subjects in Treatment C requiring significantly less time to make the nine decisions than those of the other two treatments. Similarly, the subjects in Treatment A required significantly less time than did those in Treatment B.

Analyses were performed to estimate a relationship between the number of sizes within a product group and the accuracy of choice of most economical package. Also, the number of unique size-price combinations or distinct choices within each product group was determined and related to accuracy of choice. This number of unique combinations was, in general, different from the number of sizes for each product group. This was a result of the same-sized, but different brand, packages having different prices within a product group.

Correlation analyses were performed between the number of sizes and number of correct choices and between the number of unique size-price combinations and number of correct choices for each of the three experimental conditions. For Conditions A and B, all computed coefficients were significant; for Condition C, neither of the coefficients was significant. Table 1 summarizes the results. Finally, data gathered from the 12 subjects performing the experiment in the supermarket were summarized. Table 2 presents the mean number of correct choices and the mean time spent in making the nine choices for these subjects, together with the same information for corresponding groups in the simulation.

Although analyses of variance or  $t$  tests to test the differences between the in-store subjects and the simulation subjects on the two measures were not performed because of large

TABLE 1

CORRELATION COEFFICIENTS BETWEEN NUMBER OF SIZES, NUMBER OF UNIQUE SIZE-PRICE COMBINATIONS, AND NUMBER OF ACCURATE CHOICES

Treatment	Correlation between number of sizes and correct choices	Correlation between number of unique size-price combinations and correct choices
A	.50*	.48*
B	.40*	.45*
C	.04	.03

\*  $p < .05$ .

discrepancies in sample size, it would appear nevertheless that the pattern of data obtained for the in-store subjects parallels those engaged in larger experiments.

### DISCUSSION

The results of experimentation offer support to those that favor unit pricing as a method of presenting information to consumers about weight and price of supermarket items.

The first hypothesis, that consumers are significantly more accurate in their choices of "most economical" when receiving unit-price information, was supported. A review of individual scores leads to two interesting observations. First, the variability in individual scores is considerably less for those in the unit-pricing treatment than in the other two. For each of these two treatments, scores

ranged from two correct to eight correct. In the unit-pricing condition the range was from seven to nine correct, indicating that all subjects were able to adequately process this information. This would seem to be a highly desirable end product of an information system. A second observation is that even when unit-pricing information is presented, further education on economical purchasing is needed. Only 3 of the 25 subjects accurately chose the economical package of instant potatoes, despite instruction that "most economical" could be judged in terms of information in addition to price per unit. Apparently, subjects developed a set that price per unit of net weight was always the most economical. Therefore even with unit pricing for most goods, it would seem necessary to inform consumers that they would be cognizant of factors such as strength of solution (bleach, artificial sweeteners) or one-ply or two-ply construction (tissues), when determining economy of purchase.

The computational aid did not improve the accuracy of consumer choices when compared with choices made under present supermarket methods of presenting information, supporting Hypothesis 3. This device was designed to make price-per-ounce calculations faster and more accurate for consumers. However, this calculation is only one part of the information a consumer must process before making a decision; he must still keep account of the price per unit for each package and make a judgment based on this information.

TABLE 2  
DATA COMPARING PERFORMANCE OF IN-STORE SUBJECTS AND SIMULATION SUBJECTS FOR TREATMENTS A AND B

Item	Treatment A				Treatment B			
	In-store condition <sup>a</sup>		Simulation condition <sup>b</sup>		In-store condition <sup>a</sup>		Simulation condition <sup>b</sup>	
	$\bar{X}$	$\sigma$	$\bar{X}$	$\sigma$	$\bar{X}$	$\sigma$	$\bar{X}$	$\sigma$
No. of correct choices	5.07	.82	5.72	1.31	5.34	1.29	5.96	1.57
No. of minutes to make nine choices	27.83	11.27	23.93	10.00	35.12	11.76	31.72	9.57

<sup>a</sup>  $n = 6$ ,  
<sup>b</sup>  $n = 25$ .

This apparently is a difficult task, leading to errors of choice. A possibility for error is also introduced in the manipulatory aspects of the device. If the consumer inadvertently matches the wrong numbers on the device, he will naturally receive the wrong information. There would seem to be numerous possible occasions to commit such inadvertent mistakes, including misreading information on the package, misreading the entries on the aid, and accidentally moving the wheel on the aid.

The second hypothesis, that significantly less time is required for consumers to choose the most economical package when unit-pricing information is presented, was also supported.

Inspection of individual scores indicate very little variability in the time needed to make the nine decisions under unit pricing. For this treatment, all subjects completed the task in from 2 to 4 minutes. For Treatment A, the range was from 11 to 56 minutes and for Treatment B, from 23 to 60 minutes. This is an important observation. Speed in processing information would also seem to be a desirable end product of a pricing system.

Hypothesis 4, that there is no difference in time required to arrive at decisions under the two conditions other than the unit-pricing condition, was not supported. Obviously, the time required to manipulate the computational device for each calculation added greatly to the total shopping time.

Hypothesis 5, that the number of sizes of packages within a product group is inversely correlated to accuracy of choice for Treatments A and B, but not correlated for Treatment C, was not supported. For Treatments A and B, significant positive correlations were found between the number of correct choices and the number of sizes and also between the

number of correct choices and the number of unique size-price combinations. These findings are significant in that one thrust of present governmental work to aid consumers in making price comparisons is to reduce the number of sizes within a product group. Correlations found in this study, of course, are based on limited data and therefore are not definitive; but they indicate that such a strategy may not be appropriate.

A few observations on the representativeness of the simulation seem to be appropriate. For this experiment, there are two points of comparison. The first is comparing the performance of the subjects in Treatment A with that of subjects in the previously reported study. In the study conducted by Friedman (1966), subjects performed the experimental task in actual supermarkets, with an accuracy rate of 57%. The accuracy rate for comparable subjects in the present experimentation was 63%, suggesting the comparability of the simulation with higher fidelity experimentation. A second comparison can be made between the performance of the simulation subjects and those that made their choices in the store (Table 2). For both treatments, performance differences were small, further supporting the contention of representativeness.

#### REFERENCES

- COILAN, S. E., Packaging law is on book, but ills it aimed to cure are still troublesome. *Advertising Age*, 1969, 40, 10.
- FRIEDMAN, M. P. Consumer confusion in the selection of supermarket products. *Journal of Applied Psychology*, 1966, 50, 529-534.
- Grocers moan, but New York moves on unit prices. *Advertising Age*, 1969, 40, 3.
- Mrs. Knauer twits commerce, says it fails to cut package proliferation. *Advertising Age*, 1969, 40, 1.

(Received July 21, 1971)

## SHORT NOTES

# A TEST OF THE NEED GRATIFICATION THEORY OF JOB SATISFACTION<sup>1</sup>

JAMES D. NEELEY, JR.<sup>2</sup>

*Cornell University*

Wolf (1970) has advanced the need gratification theory of job satisfaction/dissatisfaction that emphasizes the moderating influence of psychological needs on the relationship between job elements and satisfaction. This study tests two need gratification hypotheses using 75 nonacademic college employees and assessing psychological needs by projective methods. Neither of the original hypotheses were supported. It is suggested that need gratification theory be expanded to include situation variables.

Need gratification theory (Wolf, 1970) has been proposed as an alternative to the two-factor theory of job satisfaction (Herzberg, Mausner, & Snyderman, 1959). Two-factor theory states that job content elements (e.g., achievement, advancement, recognition, and responsibility) are the major source of satisfaction while job context elements (e.g., company policy and administration, working conditions, and relations with other employees) are the main source of dissatisfaction. Need gratification theory introduces the consideration of the individual's psychological needs (Maslow, 1954) and their influence on the relationship between job elements and satisfaction.

The present research tests two key hypotheses of need gratification theory: (a) persons having lower level needs obtain satisfaction and dissatisfaction primarily from context elements, and (b) persons whose lower level needs are conditionally gratified and whose higher level needs therefore are active obtain satisfaction primarily from content elements and dissatisfaction from both content and context elements.

### METHOD

#### Subjects

The subjects were a stratified random sample of the nonacademic employees of a small, rural college. The sample totaled 89 (41 male, 48 female), including 17 unskilled (10 male custodians, 7 female

<sup>1</sup> This article was based on a master's thesis submitted to the Department of Psychology, Cornell University, 1970. The research was supported in part by NIGMS Training Grant GM01941-01. The author is indebted to Henry A. Alker, the author's thesis committee chairman, to Lawrence K. Williams, who also served on the thesis committee, and to Martha Feustel, Joan Gang, and John Turney.

<sup>2</sup> Requests for reprints should be sent to the author, Department of Psychology, Morrill Hall, Cornell University, Ithaca, New York, 14850.

maids), 21 skilled (19 male tradesmen and supervisors, 2 female supervisors), 32 clerical (3 male clerks, 29 female typists and secretaries), and 19 administrative-professional (9 male and 10 female librarians, managers, assistant directors and directors). The subjects varied in age from about 18 to 63 years ( $M = 37.13$ ,  $SD = 13.82$ ) and in length of service from six months to 15 years ( $M = 3.79$ ,  $SD = 3.13$ ). About 125 employees were originally invited to voluntarily participate in the study.

#### Measures

A questionnaire was administered to the subjects in group sessions. The following measures were included:

*Need hierarchy.* This variable is defined by several patterns of n Achievement, n Affiliation, and n Power. It distinguishes four types of individuals: those whose esteem needs are predominant (i.e., high n Achievement and/or high n Power combined with low n Affiliation), those whose esteem and affiliation needs are predominant (i.e., high n Achievement and/or high n Power combined with high n Affiliation), those whose affiliation needs are predominant (i.e., high n Affiliation combined with low n Achievement and low n Power), and those whose safety needs are predominant (i.e., low on all three needs).

The groups of needs defining each of these four types of individuals are given by Maslow's (1954) theory which states that achievement and power are esteem needs, and that affiliation is a love and belongingness need. Persons scoring low on all three needs might therefore have either self-actualization needs predominant (i.e., esteem and affiliation needs satisfied) or safety needs predominant (i.e., esteem and affiliation needs not yet active). However, the scarcity of self-actualizing persons in the general population argues for the latter assumption, which was adopted in this study. A person who has strong achievement and/or power in addition to affiliation needs must be between the esteem and the love and belongingness categories in the hierarchy, since some higher needs have emerged and yet lower ones have not been fully satisfied. Since the need hierarchy theory does not require that all esteem level persons

have both achievement and power needs active simultaneously, persons scoring high on just one of these needs were not distinguished in this study from persons scoring high on both. It should be noted that this measure of the need hierarchy was composed of only three scored needs, whereas Maslow's original conception included many more. The percentage distribution for the four types of needs was as follows: 35% esteem, 32% affiliation-esteem, 17% affiliation, and 16% safety.

The three Murray needs were measured from TAT stories written by the subjects in response to stimulus cards selected by Veroff, Atkinson, Feld, and Gurin (1960) to be most suitable for a sample of a cross section of the general population. Five of the six cards were different for men than for women. The test was administered and blindly scored by the investigator following the standard procedures presented in Atkinson (1958). A scoring reliability check provided by an assistant on a random sample of 108 stories given by 18 subjects yielded a tetrachoric  $r$  of .87 for n Achievement, .72 for n Affiliation, and .75 for n Power. Both the investigator and the assistant had previously and independently obtained high reliabilities with the scoring manuals in Atkinson (1958). Following the procedure given by Veroff et al. (1960), raw motive scores were adjusted to be independent of the length of the TAT protocol. None of the intercorrelations of the three needs were significant.

**Critical-incident stories.** Two critical incident stories, written by the subjects, were used to determine which kind of job elements made them satisfied or dissatisfied. The directions for writing the stories followed those in Herzberg et al. (1959). The stories were scored blindly by the investigator using the scoring guide in Herzberg et al. (1959). The content of the stories was distinguished only as predominantly content element- or context element-related. A scoring reliability check by an assistant on a random sample of 36 scorable stories yielded a reliability of .94 (tetrachoric  $r$ ). Protocols were complete enough to be scored for 75 (84%) of the subjects.

Of interest in this study was the pattern of content and context element themes in the individual's critical incident stories. For any individual, this pattern will be one of the following four: (a) Content-Context; the story about a time when the subject felt good about (was satisfied with) his job was predominantly content-related, while the story about a time when he felt bad (was dissatisfied) was predominantly context-related, (b) Content-Content; both stories were predominantly content-related, (c) Context-Context; both stories were predominantly context-related, or (d) Context-Content; the story about a time when the subject felt good about his job was predominantly context-related, while the story about a time when he felt bad was predominantly content-related. Need gratification theory predicts the Content-Context and Content-Content patterns for those individuals with higher level needs, and the Context-Context pattern for

TABLE 1  
JOB ELEMENT PATTERN FREQUENCIES FOR HIERARCHICAL NEED GROUPS

Predominant need in hierarchy	Job element patterns	
	Content-context or content-content	Context-context
Safety	11	0
Affiliation	11	2
Affiliation-esteem	16	5
Esteem	22	4

Note.  $N = 75$ . Four of these gave the Context-Content pattern, which was omitted from this study.

those with lower level needs. The Context-Content pattern is irrelevant to need gratification theory and was not considered in this study.

The percentage distribution for these patterns was 43% Content-Context, 37% Content-Content, 15% Context-Context, and 5% Context-Content. That the Context-Context pattern represented only 15% of the cases may have precluded a test of the hypothesis that persons having lower level needs obtain both their satisfaction and dissatisfaction primarily from context elements. However, 33% of the sample did have either safety or affiliation (lower level) needs, and so about one third of the sample would have been expected to give the Context-Context pattern. The sample was representative of the employed population, and it seems doubtful that it was biased against the appearance of this pattern.

## RESULTS AND DISCUSSION

Both need gratification theory hypotheses were tested simultaneously by performing a chi-square test on the data presented in Table 1. The null hypothesis was not rejected ( $\chi^2 = 3.13$ ,  $df = 3$ ,  $p = .38$ ), and it was concluded that differences in psychological needs were not associated with differences in the kind of job elements that were satisfying/dissatisfying. Of course, the negative results of this overall test obviated separate tests of the two original need gratification hypotheses.

The failure of need gratification theory to be supported by this research suggests a deficiency in the formulation of that theory. Following a comprehensive review of job satisfaction research, Vroom (1964) observed that studies investigating the influence of personality variables alone are not as likely to increase our understanding as are studies which examine the simultaneous effects of both personality and situation variables on job satisfaction. This criticism may also apply to need gratification theory which in its present

form contains no provision for situation differences and their potential interaction with psychological needs. Perhaps this is a direction in which need gratification theory could be profitably developed.

## REFERENCES

- ATKINSON, J. (Ed.) *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work*. New York: Wiley, 1959.

MASLOW, A. *Motivation and personality*. New York: Harper, 1954.

VEROFF, J., ATKINSON, J., FELD, S., & GURIN, G. The use of thematic apperception to assess motivation in a nationwide interview study. *Psychological Monographs*, 1960, 74(12, Whole No. 499).

VROOM, V. *Work and motivation*. New York: Wiley, 1964.

WOLF, M. Need gratification theory: A theoretical reformulation of job satisfaction/dissatisfaction and job motivation. *Journal of Applied Psychology*, 1970, 54, 87-94.

(Received July 26, 1971)

*Journal of Applied Psychology*  
1973, Vol. 57, No. 1, 88-89

## THE PREDICTIVE VALIDITY OF PREMILITARY PERFORMANCE RATINGS BY HIGH SCHOOL PERSONNEL<sup>1</sup>

JAMES M. ORVIK<sup>2</sup>

*U. S. Navy Medical Neuropsychiatric Research Unit, San Diego, California*

Faculty members in large and small school settings rated former students as to their future performance as Marine Corps enlisted men. The ratings were evaluated against a criterion of attrition and pay grade. Validity coefficients were generally low but valid. Ratings made by males were significantly higher in validity than ratings by females. Ratings made in small school settings were more valid than those made in the large school settings. Suggestions were given for modifying the use of these ratings to increase their predictive validity.

In the present investigation high school officials were requested to rate the probable success of former pupils as Marine Corps enlisted men. Two variables, the sex of the rater, and the size of the high school, were hypothesized to moderate the predictive validity of the ratings. That is, ratings by females would have lower predictive validity than those made by males. Likewise, ratings of enlistees from larger high schools would have lower predictive validity than ratings of enlistees from smaller high schools.

## METHOD

## Subjects

The subjects<sup>3</sup> were 5,172 enlisted Marines whose status was known at the end of two years of military service. Their mean age on enlistment was 18.3 years and they had attained a mean of 10.97 years of formal education. 56.8% of the subjects were high school graduates, while 8.1% had completed fewer than 9 years of school.

<sup>1</sup> This study was conducted as part of Work Unit MF022.01.049001, under the Bureau of Medicine and Surgery, Navy Department. Opinions expressed are

## Procedure

The prediction made by school officials of the subjects' adjustment to the Corps was measured by a questionnaire sent to the school by the U. S. Navy Medical Neuropsychiatric Research Unit. That questionnaire is a modified version of one developed by Flyer (1963). It was to be completed either by the principal, a guidance counselor, or a teacher who knew the subject, and required information from school records as well as subjective opinions by the rater as to school achievement, extracurricular ac-

those of the author and are not to be construed as necessarily reflecting the official view or endorsement of the Department of the Navy.

<sup>2</sup> Requests for reprints should be sent to James M. Orvik, now at the Center for Northern Educational Research, University of Alaska, College, Alaska 99701.

<sup>3</sup> In September of 1961, the staff of the U.S. Navy Medical Neuropsychiatric Research Unit began an evaluation of the selection process for enlisted personnel in the U.S. Marine Corps. The subjects of the present study were drawn from a sample of 13,447 recruits selected over a 12-month period at the two recruit training centers, San Diego, California and Parris Island, South Carolina.

tivities, and social and family adjustment. The final item on the questionnaire required the person filling out the form to make a prediction, on an itemized 4-point scale, of subject's probable success in the Marine Corps. Response to this item constituted the primary predictor variable in the present study.

Subjects were divided into four subgroups based on (a) the sex of the person who rated him and (b) the size of the school in which he and the rater previously were associated. Rather than attempt to define small and large schools by a priori criteria, small and large schools were differentiated relative to their median size for the present sample. Thus, schools with an enrollment of more than 1,000 students were designated as large schools and those with less than 1,000 students were designated as small schools. These divisions resulted in a four-fold classification of raters: (a) female raters in small schools ( $n = 245$ ), (b) female raters in large schools ( $n = 394$ ), (c) male raters in small schools ( $n = 2,312$ ), and (d) male raters in large schools ( $n = 2,221$ ).

The predictive validity of the high school ratings was established against a dichotomous, success-non-success, performance criterion. Subjects who survived 2 full years of Marine Corps service and attained a pay grade of E-2 or higher at the end of that period were defined as successful on the performance criterion. Subjects discharged from the Marine Corps prior to 2 years of service, for any of the following reasons: unfitness, misconduct, or sentence of court-martial, or whose pay grade at the end of two years was the same as their initial enlistment pay grade (E-1) were defined as nonsuccessful.

### RESULTS

The relationships between raters' predictions and the performance criterion were estimated by biserial correlation coefficients, presented in Table 1. Significance tests<sup>4</sup> of the gross differences between the validity coefficients (a) of male versus female raters and (b) of ratings from small versus large schools shows ratings by males more valid than ratings by females ( $z = 2.30$ ,  $p < .025$ ), and ratings from small schools more valid than those from large schools ( $z = 2.02$ ,  $p < .05$ ).

However, within the four-fold array of correlation values, only the difference between the correlations for females from large schools and those for males from small schools was significant ( $z = 2.42$ ,  $p < .02$ ). All other correlations values were homogeneous and significantly different from zero.

### DISCUSSION

The results showed that ratings made by males were generally more valid than ratings made by

<sup>4</sup> The tests were made by dividing the difference between the two biserial  $r$ s (since there is no  $r$  to  $z$  transformation available for the biserial  $r$ , McNemar 1962, p. 191), by the square root of the sum of the squared standard errors of the individual biserial  $r$ s (McNemar 1962, p. 191).

TABLE 1  
BISERIAL CORRELATIONS BETWEEN THE RATERS'  
PREDICTIONS AND THE TWO-YEAR  
PERFORMANCE CRITERION

Sex of Rater	Size of School					
	Small (less than 1,000 pupils)		Large (1,000 pupils or more)		Total	
	N	r	N	r	N	r
Male	2312	.28	2221	.22	4533	.25
Female	245	.21	394	.08 <sup>a</sup>	639	.11
Total	2557	.28	2615	.20	5172	.24

<sup>a</sup>  $p > .05$ ,  $p < .01$  for all other values of  $r$ .

females, and ratings made in the small school setting were more valid than ratings made in the large school setting.

The size of an enlistee's high school is obviously beyond the practical control of the military so its direct manipulation as a variable to increase the predictive validity of rating information is impossible. However, information about high school size can be of potential value when used as a moderator variable to identify subgroups of enlistees for whom high school ratings are of maximum predictive validity. From the present data, the performance of enlistees from smaller high schools is relatively more predictable than the performance of enlistees from larger high schools.

The sex of the rater, however, is best treated by modifying the instructions issued at the time the rating data are gathered. Male raters should be designated to gather the requested information, whenever possible, with particular emphasis on the prediction of the enlistee's probable success in the Marine Corps.

### REFERENCES

- EDWARDS, A. E. *Experimental design in psychological research*. New York: Holt, Rinehart & Winston, 1963.  
 FLYER, E. S. *Prediction of unsuitability among first-term airmen from aptitude indexes, high school reference data, and basic training evaluations*. (Tech. Rep. No. PLR-TDR-63-17) Lackland Air Force Base, Texas: 6570th Personnel Research Laboratory, June 1963.  
 McNEMAR, Q. *Psychological statistics*. New York: Wiley, 1962.

(Received August 23, 1971)

## AN APPLICATION OF TECHNIQUES TO SHORTEN TESTS AND INCREASE VALIDITY<sup>1</sup>

JOHN A. FOSSUM<sup>2</sup>

*Michigan State University*

Two methods of item selection against an external criterion were used to build two short tests for selecting programming and computer maintenance students. The methods were: (a) sequential accretion of items so that at each iteration the item selected is the one leading to the largest increase in the correlation between the test and the criterion and (b) accretion of items in order of their declining item point biserial correlations with the criterion. There was no significant difference in the validity of tests built using either method. Both methods produced tests with cross-valid coefficients higher than the validity of the item pool and both were reasonably resistant to shrinkage.

Two of the most important problems faced by industrial test users are low test validities and the long-time periods necessary to administer some tests. Both problems can be ameliorated, but the remedies are not usually independent. For example, when testing time is sacrificed, less information is available to use in the enhancement of validity.

This study reports the results of an attempt to attack both problems simultaneously for the entrance tests used by a large private computer technology training organization operating several U.S. and Canadian schools. Two curricula were offered, programming and computer maintenance. Tests were to be developed to predict the probability of an applicant passing the course for which he was applying. Each test was to take about 15 minutes to administer and was to contain 25 items, enough to appear face valid to applicants. For each test, the criterion against which items were to be selected was final course average. This average consisted of a weighted sum of weekly quiz grades and practice problems. All schools in the organization used standardized curricula, quizzes, and problems.

Several methods for constructing tests to increase validity or shorten test taking time have been suggested (cf. Anastasi, 1953; Darlington & Bishop, 1966; Gulliksen, 1950). Most of the techniques consider item validity and item reliability estimates in the item selection strategy. Some make use of interitem correlations (e.g.,

<sup>1</sup> The author would like to thank Professors Frank Schmidt, Michael Moore, and Dr. Walter Tornow for their helpful comments on earlier drafts of this article.

<sup>2</sup> Requests for reprints should be sent to John A. Fossum, now with the Department of Business Administration, University of Wyoming, Box 3275, University Station, Laramie, Wyoming 82070.

Darlington & Bishop, 1966). One method which attempts to maximize validity without consideration for internal consistency reliability is the sequential item nominator technique developed and programmed by Moonan and Pooch (1966). Their algorithm builds tests by selecting, at iteration one, the item that has the largest point biserial correlation with the criterion, and continues by adding items, one at a time, by determining which item has the highest semipartial correlation between item and criterion, holding the test of previously selected items constant. Using this method, tests of high validity or high reliability can be constructed depending on whether an external or internal criterion is selected.

In this study, tests for the two curricula were constructed using the sequential item nominator technique and a technique which included items in order of their declining item-criterion correlations without considering item-test correlations. This comparison was made to determine whether the additional computational complexity necessary for the sequential item nominator resulted in any significant increase in the validity of tests constructed using that method in place of the simpler method.

### METHOD

Students from three schools in each curriculum were given two different pools of items during the first week of class. At the end of the course, criterion information was collected for each student who had completed the course or had discontinued for academic reasons. Of the total sample, programming students were randomly assigned to development and cross-validation samples. Computer maintenance students from two of the schools were assigned to the development group, while students from the third school were assigned to the cross-validation group because of a lag in obtaining information from

that school. Sizes of the samples, characteristics of the item pools, and the type of items selected by each method are given in Table 1.

Validation was done concurrently with item selection since the test was built against an external criterion. Cross-validation was accomplished by summing the scores for the selected items for each subject in the cross-validation samples and correlating the total scores with each student's final average.

## RESULTS

Table 2 shows the values of three statistics at various stages of test development. For both tests developed by either method, the validity coefficient peaked before the arbitrarily selected 25-item length was reached. The sequential item nominator technique built tests with higher validities for both curricula. At the same time, the average item intercorrelation for both tests was lower using the sequential method. When average item-criterion correlations are compared, the sequential method mean was .18 for the programmer test and .20 for the computer maintenance test while the means were .22 and .24, respectively, for the declining item validity method.

Both item selection methods substantially increased validity coefficients in both curricula. The sequential technique developed tests of higher validities but was subject to greater shrinkage in the programmer sample. For the computer maintenance samples, the cross-valid correlation was slightly higher than the validity coefficient for both samples. While this is unusual, it should be remembered that students were not randomly assigned to the cross-validation sample for this test. Table 3 demonstrates that both techniques significantly increased validity and shortened test-taking time for the computer maintenance test, and shortened testing time at no detriment to validity for the programmer test.

TABLE 1  
ITEMS POOLS AND VALIDATION SAMPLES

Samples and pools	Computer maintenance	Programming
Development sample	75	209
Cross-validation sample	24	106
Total item pool	77	92
Number sequences	15	32
Verbal analogies	30	40
Formula derivations	12	20
Word problems	20	—
Items selected by SIN	25	25
Number sequences	7	7
Verbal analogies	10	10
Formula derivations	2	8
Word problems	6	—
Items selected by DIV	25	25
Number sequences	5	7
Verbal analogies	8	8
Formula derivations	4	10
Word problems	8	—
Number of overlapping items	16	16

Note. SIN = sequential item nominator, and DIV = declining item validity.

## DISCUSSION

If item intercorrelations are .00, selection of items in terms of declining item validity is as good as any other method. However, almost any item pool contains items which are intercorrelated. The sequential item nominator method is an improvement over the single item method whenever there are item intercorrelations, since these are considered to the extent to which they correlate with the previously selected items comprising a test. Both of the examples developed in

TABLE 2  
COMPARATIVE STATISTICS BY TEST AND TECHNIQUE

Test length	Computer maintenance test								Programmer test							
	$r_{ie}^a$		$r_{xy}$		$\bar{r}_{ij}$		$r_{ie}^a$		$r_{xy}$		$\bar{r}_{ij}$					
	SEQ	DIV	SEQ	DIV	SEQ	DIV	SEQ	DIV	SEQ	DIV	SEQ	DIV	SEQ	DIV	SEQ	DIV
5	.32	.31	.63	.60	.08	.13	.14	.27	.53	.50	.08	.20				
10	.05	.25	.71	.66	.09	.12	.17	.23	.58	.50	.08	.21				
15	.34	.21	.71	.66	.11	.14	.04	.19	.60	.52	.08	.18				
20	.22	.20	.68	.65	.12	.14	.13	.17	.62	.54	.09	.15				
25	.21	.16	.67	.61	.11	.14	.14	.16	.62	.54	.09	.13				

Note. SEQ = sequential item nominator, and DIV = declining item validity.  
Item-criterion correlation for item added at that iteration.

TABLE 3  
CROSS-VALIDATION RESULTS

Stage	Computer maintenance			Programmer		
	Item pool	SEQ	DIV	Item pool	SEQ	DIV
Validation	.33*	.67**	.61**	.40**	.62**	.54**
Cross-validation	--	.69**	.63**	--	.42**	.40**

Note. SEQ = sequential item nominator, and DIV = declining item validity.

\*  $p < .01$ .

\*\*  $p < .001$ .

this study indicate that consideration of item intercorrelations will result in the selection of items with a lower average item-criterion correlation and a lower average item intercorrelation for the developed test. If the item intercorrelation matrix is stable across samples, then the sequential method is superior to one which does not consider intercorrelations.

Several conclusions can be drawn from this

study. If item intercorrelations are low, there is little advantage in using the more complex sequential item nominator method. Both methods developed tests of higher validities than the item pool and both demonstrated reasonable resistance to shrinkage in these samples. The user must be aware that a test developed by either method is specific to the criterion against which it is developed. If this criterion is not stable over time, the test will have limited usefulness.

#### REFERENCES

- ANASTASI, A. An empirical study of the applicability of sequential analysis to item selection. *Educational and Psychological Measurement*, 1953, 13, 3-13.
- DARLINGTON, R. B., & BISHOP, C. H. Increasing test validity by considering interitem correlations. *Journal of Applied Psychology*, 1966, 50, 322-330.
- GULLIKSEN, H. *Theory of mental tests*. New York: Wiley, 1950.
- MOONAN, W. J., & POOCH, U. W. SEQUIN: A computerized item selection procedure. (USNPRRA Research Memorandum SRM 67-8) San Diego: U.S. Naval Personnel Research Activity, 1966.

(Received August 11, 1971)

*Journal of Applied Psychology*  
1973, Vol. 57, No. 1, 92-94

## VOLUNTEER SATISFACTION WITH IN-COUNTRY TRAINING FOR THE PEACE CORPS: REANALYSES AND EXTENDED FINDINGS<sup>1</sup>

RICHARD R. JONES<sup>2</sup>

Oregon Research Institute, Eugene

W. J. BURNS

Harvard University

Reanalyses of volunteer satisfaction with training strengthened and extended an earlier conclusion that a moderate time period for in-country training was preferred to either extensive or no in-country training. Interproject differences in satisfaction with training, beyond those attributable to levels of in-country experience, seemed to suggest the need for standardizing training programs across study areas and, perhaps, across countries.

In an earlier report (Jones & Burns, 1970) Peace Corps volunteers' satisfaction with training was studied in eight Indian projects, classified according to the amount of training time volunteers spent in India—heavy, light, and no in-country training (ICT). The results showed significant differences in satisfaction with several

components of training, both between the three levels of ICT and among projects nested within levels of the ICT factor. However, these results represent only some of the actual significant differences; a subsequently detected calculation error<sup>3</sup> revealed additional findings requiring an

<sup>1</sup> This research was supported by the Peace Corps, Contract No. PC 80-1545.

<sup>2</sup> Requests for reprints should be sent to Richard R. Jones, Oregon Research Institute, P.O. Box 3196, Eugene, Oregon 97403.

<sup>3</sup> The writers are indebted to John W. Cotton for pointing out an error in calculating the degrees of freedom for the original nested analyses of variance. A comparison between Table 1 in the present report and Table 3 in the earlier article shows the degrees of freedom for projects should be 5, not 21, and the degrees of freedom for error should be 240, not 224.

extension of the original conclusions to other components of training. This brief report provides both a correction of earlier analyses and a discussion of heretofore unreported significant findings.

For details of procedure, the reader is referred to the original report. In brief, the study was based on 248 Peace Corps volunteers assigned to eight different projects: Four heavy ICT projects (6 to 9 weeks of US training [UST] and 4 to 6 weeks of ICT); two light ICT projects (11 weeks UST and 2 weeks ICT), and two no ICT projects (14 weeks UST only). A Training Evaluation Questionnaire administered to all volunteers was scored for six components plus an overall measure of satisfaction-with-training. Each of these seven measures was treated as a dependent variable in a nested analysis-of-variance design to determine the significance of differences in training satisfaction due to the amount of in-country training and to variations among projects within ICT groups. The dependent variables were measures of satisfaction with the following components: Language training, area studies, job skills training, medical training, adjustment to India, understanding PC goals, and the summary measure of general satisfaction (obtained by averaging the six component scores).

## RESULTS AND DISCUSSION

The ANOVA findings are given in Table 1 in the present paper; the group means are presented in Table 4 in the earlier report. The reanalyzed data yielded six significant differences (in addition to the original five) among either projects or ICT groups for the set of seven dependent variables. For the ICT factor, the corrected results show that significant differences were obtained for all variables except area studies and adjustment to India. For all five significant variables, the light ICT group showed higher average satisfaction than either the heavy ICT or no ICT groups. Further, except for one tie (medical training), the means for all significant variables were lower for the no ICT group than for the heavy ICT group. Our original conclusion—that light ICT seems to be preferred training arrangement—is strengthened by the corrected analyses. Table 4 in the original report shows the magnitudes of the differences between the means of these satisfaction measures for the ICT groups.

For the project factor, only the language-training component failed to show significant differences in average satisfaction among projects. Compared with the other training components, language instruction is probably the most stan-

TABLE 1  
ANOVAs FOR SEVEN TRAINING  
SATISFACTION COMPONENTS

Language training					
Source	SS	df	MS	F	p*
ICT	15.31	2	7.65	4.27	.05
Project	12.71	5	2.54	1.42	ns
Error	430.06	240	1.79		
Area studies					
ICT	4.02	2	2.01	1.14	ns
Project	61.52	5	12.30	6.95	.001
Error	424.74	240	1.77		
Job skills					
ICT	56.25	2	28.12	11.25	.001
Project	110.79	5	22.16	8.86	.001
Error	598.80	240	2.50		
Medical training					
ICT	40.41	2	20.21	9.76	.001
Project	268.02	5	53.60	25.89	.001
Error	497.79	240	2.07		
Adjustment to India					
ICT	9.25	2	4.63	1.51	ns
Project	42.79	5	8.56	2.79	.05
Error	736.42	240	3.07		
Understanding PC goals					
ICT	19.97	2	9.99	3.23	.05
Project	36.06	5	7.21	2.33	.05
Error	741.37	240	3.09		
General satisfaction					
ICT	9.17	2	4.58	3.09	.05
Project	38.58	5	7.72	5.22	.001
Error	354.67	240	1.48		

\* For testing the ICT factor,  $F(2, 240)$  equals 2.99 at the .05 level and 4.60 at the .01 level. For the project factor,  $F(5, 240)$  equals 2.21 at the .05 level and 3.02 at the .01 level.

dardized, least project-specific component of training. No doubt, fewer project differences arise from language-training variables—Instructors, course materials, teaching methods—than from the other training components. If this interpretation is correct and these findings are

generalizable to other PC countries, qualitative differences among projects (D. Jones, 1968) might be reduced by standardizing instructional procedures in other training-component areas.

### CONCLUSIONS

In sum, these reanalyses of the satisfaction measures support and extend the original conclusion that light ICT seems to be the optimal arrangement of US vs. in-country training, at least for this sample of Indian projects. Project differences within and across ICT levels represent an important secondary finding. Certainly, it seems clear that in spite of significant variations

attributable to ICT levels, inter-project differences in volunteer satisfaction are at least as important and point to a need for changes in programming of training activities to reduce these kinds of variability.

### REFERENCES

- JONES, D. *The making of a volunteer: A review of Peace Corps training—Summer, 1968*. Washington, D.C.: Office of Evaluation, Peace Corps, December 1968.  
JONES, R. R., & BURNS, W. J. Volunteer satisfaction with in-country training for the Peace Corps. *Journal of Applied Psychology*, 1970, 54, 533-537.

(Received July 26, 1971)

### Elimination of Early Publication Policy

At the November 5-6, 1972 meeting of the Publications Board, the Board agreed that publication needs should be met through page allocation and editorial policy rather than through the use of early publication practices. The Board noted that the production costs per page were not uniform across journals, though the charge to authors was, and that the number of early publication pages had decreased so significantly that it was apparent that the practice was no longer needed. The Board therefore voted to rescind its policy of permitting early publication in the journals. This action does not apply to articles already designated for publication before January 1974.

# JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1973 by the American Psychological Association, Inc.

April 1973

Vol. 57, No. 2

## ARTICLES

- The Relationship between Sex Role Stereotypes and Requisite Management Characteristics *Virginia Ellen Schein* 95
- Ethnic Group Differences in Relationships among Criteria of Job Performance *Alan R. Bass and John N. Turner* 101
- Predicting the Effects of Leadership Training and Experience from the Contingency Model: A Clarification *Fred E. Fiedler* 110
- Predicting the Effects of Leadership Training and Experience from the Contingency Model: Some Remaining Problems *Steven Kerr and Anne Harlan* 114
- Managerial Satisfactions and Organizational Roles: An Investigation of Porter's Need Deficiency Scales *Jeanne B. Herman and Charles L. Hulin* 118
- Equity Theory and Career Pay: A Computer Simulation Approach *Paul C. Nystrom* 125
- Internal-External Control as a Predictor of Task Effort and Satisfaction Subsequent to Failure *Howard Weiss and John Sherman* 132
- Professional Employees' Preference for Upward Mobility *Dorothy N. Harlow* 137
- Effect of Home Environment Tobacco Smoke on Family Health *Paul Cameron and Donald Robertson* 142
- The Life History Questionnaire as a Predictor of Performance in Navy Diver Training *Robert Helmreich, Roger Bakeman, and Roland Radloff* 148
- Color Versus Numeric Coding in a Keeping-Track Task: Performance under Varying Load Conditions *Jacelyn Wedell and David G. Alden* 154
- Written Information: Some Alternatives to Prose for Expressing the Outcomes of Complex Contingencies *Patricia Wright and Fraser Reid* 160
- Market Image as a Function of Consumer Group and Market Type: A Quantitative Approach *Alan Pohlman and Samuel Mudd* 167
- Effects of Human Models on Perceived Product Quality *Bernard N. Kanungo and Sam Pang* 172
- Personality and Product Use Revisited: An Exploration with the Personality Research Form *Parker M. Worthing, M. Venkatesan, and Steve Smith* 179
- Dimensions of Attitudes toward Technology *Roy D. Goldman, Bruce B. Platt, and Robert M. Kaplan* 184

## SHORT NOTES

- An Evaluation of Item-by-Item Test Administration *Cecil J. Mullins and Iris H. Massey* 188
- A Longitudinal Predictive Study of Success and Performance of Law Enforcement Officers *Stanley P. Azen, Homa M. Snibbe, and Hugh R. Montgomery* 190
- Dimensional Analysis of the Least Preferred Co-Worker Scales *William M. Fox, Walter A. Hill, and Wilson H. Guertin* 192

## LIST OF MANUSCRIPTS ACCEPTED

166

## INFORMATION FOR CONTRIBUTORS

*Style.* A professional article should possess certain characteristics: (a) conciseness and an apparent respect for reader time; (b) unambiguous and simple vocabulary with technical and erudite words used only when simpler ones would obviously be inadequate; (c) conformity to accepted technical style in tables, terminology, and references; (d) conclusions that are clearly related to the evidence presented. The reader should be led, step by step, from a statement of problem or purpose, through analysis of evidence, to conclusions and implications. Authors are encouraged to consult the little book entitled *The Elements of Style* by W. Strunk, Jr. and E. B. White (New York, Macmillan, 1962).

*Criteria for evaluation.* Manuscripts will be evaluated on the basis of several criteria, including: (a) significance in contributing new knowledge to the field, (b) technical adequacy, (c) appropriateness for the *Journal of Applied Psychology*, and (d) clarity of presentation.

*Format.* Manuscripts must be prepared in the format described in the *Publication Manual of the American Psychological Association* (1967 Revision), obtainable for \$1.50 from the American Psychological Association. Articles not prepared in this manner cannot be reviewed. Special attention should be given to the section on typing the manuscript (p. 48) and to the sections on tables, figures, and references. Note that *all copy must be double spaced*, including references, title, figure captions, etc. The senior author's last name should appear on each page in the upper left-hand corner except when blind review is required. In reference lists, give journal titles in full; do not abbreviate.

*Optional blind review.* Blind reviewing may be obtained if specifically requested at the time a manuscript is first submitted. In such cases, the author's name and affiliation should appear only on a separate title page, which must be included with each copy of the manuscript. Footnotes containing information pertaining to the identity of the author or his affiliation should be on separate pages.

*Copies.* All manuscripts must be submitted in triplicate. One copy should be the original typed copy, the other two, clear carbon copies or photo reproductions, only. Authors should check the final typing carefully and retain a copy of the manuscript as a precaution against loss in the mail.

*Length.* Before preparing a manuscript, the author should check several recent issues to get an idea of the approximate length of regular articles published in the *Journal of Applied Psychology*. (One printed page equals roughly three double-spaced typewritten manuscript pages.) A few longer articles of special significance may be printed from time to time as monographs. Occasionally the *Journal* will have a section of "Short Notes" featuring brief reports on studies which make some methodological contribution or constitute an important replication.

*Abstracts.* Each copy of the manuscript must be accompanied by an abstract of 100-120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

*Figures.* The original drawing of a figure, or an 8 X 10-in. glossy print of the drawing, is needed for publication and must be submitted with the original typed copy of the manuscript. Duplicate copies of the figure may be photographic or pencil-drawn copies. Figures must be hand-lettered; typewritten lettering is not acceptable.

*Reprints.* No gratis reprints are supplied. Reprints may be ordered from the printer prior to publication.

*Supplementary material.* Supplementary materials formerly deposited with the National Auxiliary Publications Service need no longer be submitted with the manuscript. Authors should keep supporting and raw data for at least 5 yr. after publication of their article, and if applicable, offer them to the reader upon request.

## THE RELATIONSHIP BETWEEN SEX ROLE STEREOTYPES AND REQUISITE MANAGEMENT CHARACTERISTICS

VIRGINIA ELLEN SCHEIN<sup>1</sup>

*Life Office Management Association, New York*

Three hundred male middle managers rated either women in general, men in general, or successful middle managers on 92 descriptive terms. The results confirmed the hypothesis that successful middle managers are perceived to possess characteristics, attitudes, and temperaments more commonly ascribed to men in general than to women in general. There was a significant resemblance between the mean ratings of men and managers, whereas there was no resemblance between women and managers. Examination of mean rating differences among women, men, and managers on each of the items disclosed some requisite management characteristics which were not synonymous with the masculine sex role stereotype. Implications of the demonstrated relationship for organizational behaviors are discussed.

Although women make up 38% of the work force (Koontz, 1971), the proportion of women who occupy managerial and executive positions is markedly small. One extensive survey of industrial organizations (*Women in the Work Force*, 1970) revealed that 87% of the companies surveyed had 5% or fewer women in middle management and above.

According to Orth and Jacobs (1971), one reason for the limited number of women managers and executives is that "... traditional male attitudes toward women at the professional and managerial levels continue to block change [p. 140]." Bowman, Worthy, and Greyser (1965) found that of 1,000 male executives surveyed, 41% expressed mildly unfavorable to strongly unfavorable attitudes toward women in management. This negative reaction to women in management suggests that sex role stereotypes may be inhibiting women from advancing in the managerial work force.

The existence of sex role stereotypes has been documented by numerous researchers (Anastasi & Foley, 1949; Maccoby, 1966; Wylie, 1961). For example, Rosenkrantz, Vogel, Bee, Broverman, and Broverman (1968) found that among male and female college students, men were perceived as more

aggressive and independent than women, whereas women were seen as more tactful, gentle, and quiet than men. In addition, these researchers found that the self-concepts of men and women were very similar to their respective stereotypes.

One way in which sex role stereotypes may impede the progress of women is through the creation of occupational sex typing. According to Merton, "... occupations can be described as 'sex-typed' when a large majority of those in them are of one sex and when there is an associated normative expectation that this is how it should be [Epstein, 1970, p. 152]." Judging from the high ratio of men to women in managerial positions and the informal belief that this is how it should be, the managerial job can be classified as a masculine occupation. If so, then the managerial position would seem to require personal attributes often thought to be more characteristic of men than women. Basil (cited by Brenner, 1970), using a nationwide sample of present managers, found that the four personal characteristics rated as most important for an upper management position were seen as more likely to be possessed by men than women. Thus, in general, sex role stereotypes may effectuate the perception of women as being less qualified than men for high-level management positions.

Also, sex role stereotypes may deter women from striving to succeed in managerial positions. In a theory of work behavior, Korman (1970) maintains that "... individuals will

<sup>1</sup> I would like to thank John C. Sherman for his assistance with the statistical analyses.

Requests for reprints should be sent to Virginia Ellen Schein who is now with Personnel Research, Metropolitan Life Insurance Company, 1 Madison Avenue, New York, New York 10010.

engage in and find satisfying those behavioral roles which will maximize their sense of cognitive balance or consistency [p. 32]." If a woman's self-image incorporates aspects of the stereotypical feminine role, she may be less inclined to acquire the job characteristics or engage in the job behaviors associated with the masculine managerial position since such characteristics and behaviors are inconsistent with her self-image.

Despite the apparent influence of stereotypical attitudes on the selection, placement and promotion of women, there is a dearth of studies that analyze the operation of sex role stereotypes within organizations. Although stereotypical masculine characteristics have been found to be more socially desirable (Rosenkrantz et al., 1968) and more similar to the characteristics of the healthy adult (Broverman, Broverman, Clarkson, Rosenkrantz, & Vogel, 1970) than stereotypical feminine characteristics, Schein (1971) found a paucity of studies dealing with psychological barriers, such as sex role stereotyping, that prevent women from achieving in the work force.

Since there have been no empirical studies except for Basil's demonstrating the existence of a relationship between sex role stereotypes and the perceived requisite personal characteristics for the middle management position, the purpose of the present study was to examine this association. Specifically, it was hypothesized that successful middle managers are perceived to possess those characteristics, attitudes and temperaments more commonly ascribed to men in general than to women in general. Bowman et al. found that male acceptance of women managers increases with the age of the respondent. Therefore, it was also hypothesized that the association between sex role stereotypes and requisite management characteristics would be less strong among older managers than among younger ones.

#### METHOD

##### *Sample*

The sample was composed of 300 middle line male managers of various departments within nine insurance companies located throughout the United States. Their ages ranged from 24 to 64, with a median of 43 years, their years of experience as managers, from 1 to 40 years with the median being 10 years.

#### *Measurement Instrument*

In order to define both the sex role stereotypes and the characteristics of successful middle managers, three forms of a Descriptive Index were developed. All three forms contained the same descriptive terms and instructions, except that one form asked for a description of women in general (Women), one for a description of men in general (Men) and one for a description of successful middle managers (Managers).

In developing the Descriptive Index, 131 items that differentially described males and females were garnered from studies by Basil (In Brenner, 1970), Bennett and Cohen (1959), Brim (1958), and Rosenkrantz et al. (1968). Using these items, a preliminary form of the Descriptive Index was administered to 24 male and female college students. Half of the subjects were given the Women form and half the Men form. In order to maximize the differences in the descriptions of Women and Men, an analysis of all the means and standard deviations was performed and an item was eliminated if (a) its mean descriptive rating was the same for both Women and Men, (b) it was judged by the experimenter and a staff assistant independently to be similar in meaning to one or more other items but it had a smaller mean difference between descriptions of Women and Men, or (c) its variability on both forms was significantly greater than the overall mean variability.

The final form of the Descriptive Index contained 92 adjectives and descriptive terms. The instructions on the three forms of the Index were as follows:

5

On the following pages you will find a series of descriptive terms commonly used to characterize people in general. Some of these terms are positive in connotation, others are negative, and some are neither very positive nor very negative.

We would like you to use this list to tell us what you think (women in general, men in general, or successful middle managers) are like. In making your judgments, it may be helpful to imagine that you are about to meet a person for the first time and the only thing you know in advance is that the person is (an adult female, an adult male, or a successful middle manager). Please rate each word or phrase in terms of how characteristic it is of (women in general, men in general, or successful middle managers).

The ratings of the descriptive terms were made according to a 5-point scale, ranging from 1 (not characteristic) to 5 (characteristic) with a neutral rating of 3 (neither characteristic nor uncharacteristic).

#### *Procedure*

Within each company, a representative with research experience randomly distributed an equal number of the three forms of the Index to male managers with a salary range of approximately \$12,000 to \$30,000 and a minimum of one year of experience at the managerial level.

Each manager received only one form of the Index. The cover letter to the participants stated that the researcher was ". . . engaged in the establishment of a Descriptive Index to be used for management development" and informed the participants that "since various forms of the questionnaire are being distributed within your company, high quality research results can only be obtained if you do not discuss your questionnaire or responses to it with anyone in your company." The questionnaires were returned in individually sealed envelopes.

Of the total number of Descriptive Indexes distributed, 76.62% or 354 out of 462 were returned. The return rates for the various forms of the Index were as follows: Women, 76.62%; Men, 77.27%; and Managers, 75.97%. The usable number of questionnaires was reduced to 300 (88 Women, 107 Men, and 105 Managers). Questionnaires were eliminated if (a) demographic data, such as age and sex, were not indicated or (b) the questionnaires were completed by females. Of the latter, 17 out of 26 were Women forms, which accounts for the lower number of usable Women questionnaires.

### RESULTS

The degree of resemblance between the descriptions of Men and Managers and between Women and Managers was determined by computing intraclass correlation coefficients ( $r'$ ) from two randomized groups analyses of variance (see Hays, 1963, p. 424). The classes (or groups) were the 92 descriptive items. In the first analysis, the scores within each class were the mean item ratings of Men and Managers, while in the second analysis, they were the mean item ratings of Women and Managers. According to Hays, the larger the value of  $r'$ , the more similar do observations in the same class tend to be relative to observations in different classes. Thus, the smaller the within item variability, rela-

TABLE 1

#### ANALYSES OF VARIANCE OF MEAN ITEM RATINGS AND INTRACLASS COEFFICIENTS

Source	df	MS	F	$r'$
Men and managers				
Between items	91	1.27	4.23*	.62
Within items	92	.30		
Women and managers				
Between items	91	.89	1.13	.06
Within items	92	.79		

\*  $p < .01$ .

TABLE 2  
INTRACLASS COEFFICIENTS WITHIN  
THREE AGE LEVELS

Age level	Intraclass coefficients	
	Men and managers	Women and managers
24-39 ( $n = 113$ )	.60**	.01
40-48 ( $n = 95$ )	.64**	.00
49 and above ( $n = 92$ )	.60**	.16*

\*  $p < .05$ .\*\*  $p < .01$ .

tive to the between item variability, the greater the similarity between the mean item ratings of either Men and Managers or Women and Managers.

According to Table 1, which presents the results of the analyses of variance and the intraclass correlation coefficients, there was a large and significant resemblance between the ratings of Men and Managers ( $r' = .62$ ) whereas there was a near zero, nonsignificant resemblance between the ratings of Women and Managers ( $r' = .06$ ), thereby confirming the hypothesis that Managers are perceived to possess characteristics more commonly ascribed to Men than to Women.

To determine if age moderates the relationship, the total sample was divided into three age levels, with an approximately equal number of subjects distributed within each age level and within each Women, Men, and Manager group. Intraclass correlations between the mean ratings of Men and Managers and between Women and Managers were computed within each of the three age levels. According to the results, as shown in Table 2, the main hypothesis is less strongly supported among subjects 49 years and above than among younger subjects. Within all three age levels, there was a significant resemblance between the mean ratings of Men and Managers. Among subjects 24 to 39 years and those 40 to 48 years, there was no resemblance between Women and Managers; however, among subjects 49 years and above there was a small but significant resemblance between the ratings of Women and those of Managers.

In addition to intraclass correlation coefficients, Pearson product moment correlation coefficients were computed in order to deter-

mine the linear relationships between the mean ratings among the three groups. According to the results, there was a significant correlation ( $r = .81$ ,  $p < .01$ ) between the mean ratings of Men and Managers, but the correlation between the mean ratings of Women and Managers was not significant ( $r = .10$ ). Within all three age levels the  $r$  between Men and Managers was significant at the .01 level ( $r_1 = .77$ ;  $r_2 = .80$ ;  $r_3 = .79$ ). Within the two younger groups the correlation between Women and Managers was not significant ( $r_1 = .04$ ;  $r_2 = .05$ ); however, there was a significant correlation between the mean ratings of Women and Managers among subjects 49 years and above ( $r_3 = .23$ ,  $p < .05$ ).

Although the determination of the degree of resemblance between the mean ratings of Men and Managers and the degree of resemblance between the mean ratings of Women and Managers was considered to be the primary test of the hypothesis, an exploratory examination of the specific descriptive items on which Women or Men were perceived as similar to or different from Managers was also carried out so as to obtain a better understanding of the relationship. For each of the 92 items a  $3 \times 3$  factorial analysis of variance, incorporating the three groups (Women, Men, and Managers) and the three age levels, was performed. According to the results, there was a significant group effect for 86 of the 92 items. An alpha level of .0005 was used as the criterion of significance; therefore, the probability of obtaining one or

more spuriously significant  $F$  ratios was .045. There were no significant age effects, nor were there any significant age  $\times$  group interactions.

For each of the 86 items displaying a significant group effect, Duncan's multiple range test for unequal  $n$ 's (see Kramer, 1956) was used to determine the significance of the difference ( $\alpha = .01$ ) between the mean ratings of Men and Managers, Women and Managers, and Men and Women. The results revealed that on 60 of these 86 items, ratings of Managers were more similar to Men than to ratings of Women; for 8 of the 86 items the ratings of Managers were more similar to those of Women than to Men; and for the remaining 18 items with significant group  $F$  ratios there was no relationship between sex role stereotypes and perceptions of managerial characteristics—both the mean ratings of Women and Men were significantly different from those of Managers, but there were no significant differences between the mean ratings of Women and Men.<sup>2</sup>

Items representative of the first outcome category, in which Managers were more similar to Men than to Women, were as follows: Emotionally Stable; Aggressive; Leadership Ability; Self-Reliant; (not) Uncertain; Vigorous; Desires Responsibility; (not) Frivolous; Objective; Well Informed and Direct. These items were judged to be representative of the total group of 60 items by three advanced psychology students unfamiliar with the aims of the study. Table 3 presents the items in the latter two outcome categories, in which the predicted direction of mean differences did not occur.<sup>3</sup>

TABLE 3  
ITEMS DISPLAYING LACK OF SIMILARITY  
BETWEEN MANAGERS AND MEN

Category	Item
Managers more similar to women than to men	Understanding Helpful Sophisticated Aware of feelings of others Intuitive Neat (Not) Vulgar Humanitarian Values
Sex role stereotypes not related to management characteristics	Competent Tactful Creative Courteous (Not) Exhibitionist (Not) Devious (Not) Deceitful (Not) Strong Need for Social Acceptance (Not) Desire to Avoid Controversy (Not) Dawdler and Procrastinator (Not) Desire for Friendship Intelligent Persistent Curious (Not) Quarrelsome (Not) Hasty (Not) Bitter (Not) Selfish

<sup>2</sup> Since the 92 items are undoubtedly intercorrelated, the number of significant item differences within each of the three outcome categories should not be viewed as a test of the hypothesis. The  $N$  within each of the Women, Men, and Manager groups approximated the number of items, thereby precluding a factor analysis within groups, and a factor analysis combining the responses to the three different forms of the Descriptive Index would be misleading due to the possibility of differing factor structures within the three stimulus groups (see Nunnally, 1967).

<sup>3</sup> A complete list of the items and Women, Men, and Manager mean ratings within the three outcome categories is available upon request from the author (see address in Footnote 1).

## DISCUSSION

The results confirm the hypothesis that successful middle managers are perceived to possess those characteristics, attitudes and temperaments more commonly ascribed to men in general than to women in general. This association between sex role stereotypes and perceptions of requisite management characteristics seems to account, in part, for the limited number of women in management positions, thereby underscoring the need for research on the effect of these stereotypical attitudes on actual behavior, such as organizational decision making and individual job performance.

The results suggest that, all else being equal, the perceived similarity between the characteristics of successful middle managers and men in general increases the likelihood of a male rather than a female being selected for or promoted to a managerial position. In a study of hiring practices in colleges and universities, Fidell (1970), using hypothetical descriptions of young PhDs which were identical except for sex, found that the modal level of job offer was lower for women (assistant professor) than for men (associate professor). The present findings imply that similar types of discriminatory selection decisions occur in industrial settings.

To the extent that a woman's self-image incorporates the female sex role stereotype, this relationship would also seem to influence a woman's job behavior. For example, in a laboratory task study pairing high and low dominance subjects, Megargee (1969) found that where the same sex subjects or high dominance males and low dominance females were paired, the high dominance subject, regardless of sex, assumed the leadership role; however, where high dominance females were paired with low dominance males, the high dominance females did not assume the leadership role. In this particular pairing, evidently, assumption of the leadership role was inconsistent with the females' feminine self-image and, therefore, they preferred to maintain their cognitive consistency by not being leaders. Given the high degree of resemblance between the perceived requisite management characteristics and characteristics of men in general, women may suppress the exhibition

of many managerial job attributes in order to maintain their feminine self-image. Certainly, additional research is needed to determine if this relationship between sex role stereotypes and management characteristics exists among female middle managers.

Although approximately the same degree of resemblance between the characteristics of successful middle managers and those of men in general was found within all three age levels, only subjects within the 49 and above age group perceived a resemblance between the characteristics of Managers and those of Women. This finding suggests a slight reduction of the differential stereotypical perceptions of men and women among older managers. Examination of the degree of resemblance between the characteristics of Men and Women within the three age levels supported this notion. There was no significant resemblance between Women and Men within the two younger age levels ( $r'_1 = -.14$ ;  $r'_2 = .07$ ), whereas there was a significant resemblance between Women and Men among the oldest group of managers ( $r'_3 = .30$ ,  $p < .05$ ).

Certain concomitants of age, such as experience, may somewhat reduce the perceptual 'male-typing' of the managerial job. For example, experienced managers (the  $r$  between age and managerial experience was .76) probably have had more exposure to women as managers, thereby modifying some of their stereotypical perceptions of women. Perhaps more influential to their perceptions may be the changing roles of the wives and female social peers of these older managers. According to Kreps (1971), the proportion of women in the work force increases from age 16 until early 20s, then declines sharply but rises to a second peak of participation that is reached at about age 50. Older male managers may have more interaction with women for whom the role of labor force participant is more salient than that of mother-home-maker. This age effect interpretation implies that as more women become active participants in the labor force, the increased experience with working women will reduce to some extent the relationship between sex role stereotypes and requisite management characteristics among all age groups. Consequently, this psychological barrier to women

in management will be lowered, thereby affording a greater opportunity for women to enter into and advance in managerial positions.

The results disclosing certain managerial characteristics that were not synonymous with the masculine sex role stereotype indicate areas in which women presently may be more readily acceptable in and accepting of managerial positions. Examination of the items in Table 3 suggests that "employee-centered" or "consideration" behaviors, such as Understanding, Helpful, and Intuitive, are requisite managerial characteristics that are more commonly ascribed to women in general than to men in general. In certain situations, exhibition of these stereotypical feminine behaviors may be advantageous. For example, in an experimental study, Bond and Vinacke (1961) used a task that required coalition formation for success. Males tended to use exploitative strategies, while females tended to use accommodative techniques. For this particular task, females outperformed the males. Perhaps focusing more attention on the feminine characteristics that are related to managerial success will foster a climate of greater receptivity to women managers.

Turning again to Table 3, some of the perceived requisite characteristics that were not related to sex role stereotypes, such as Intelligent, Competent, and Creative, can be classified as ability or expertise factors. That expertise is perceived to be as characteristic of women as of men supports Brenner's suggestion that women can be placed in managerial positions in which expertise is an important component of authority and explains Bowman et al.'s. finding that male managers perceive more opportunity for women managers in staff than in line positions. Most of the remaining items in this outcome category appear to be socially undesirable personality traits, such as Quarrelsome, Bitter, Devious, and Deceitful. These traits were less characteristic of successful managers than of either men or women, but no difference in the possession of these traits was perceived between men and women. Here, too, accentuation of the finding that certain attributes required of successful managers may be found more or less as easily among women as men may enhance the status of women in management.

## REFERENCES

- ANASTASI, A., & FOLEY, J. P., JR. *Differential Psychology*. New York: Macmillan, 1949.
- BENNETT, E. M., & COHEN, L. R. Men and women: Personality patterns and contrasts. *Genetic Psychology Monographs*, 1959, 59, 101-155.
- BOND, J. R., & VINACKE, W. E. Coalitions in mixed-sex triads. *Sociometry*, 1961, 24, 61-75.
- BOWMAN, G. W., WORTHY, N. B., & GREYSER, S. A. Are women executives people? *Harvard Business Review*, 1965, 43, 14-16+.
- BRENNER, M. H. Management development activities for women. Paper presented at the meeting of the American Psychological Association, Miami, September 1970.
- BREIM, O. G. Family structure and sex-role learning by children: A further analysis of Helen Koch's data. *Sociometry*, 1958, 21, 1-16.
- BROVERMAN, I. K., BROVERMAN, D. M., CLARKSON, F. E., ROSENKRANTZ, P. S., & VOGEL, S. R. Sex-role stereotypes and clinical judgments of mental health. *Journal of Consulting and Clinical Psychology*, 1970, 34, 1-7.
- EPSTEIN, C. F. *Woman's place*. Berkeley: University of California Press, 1970.
- FIDELL, L. S. Empirical verification of sex discrimination in hiring practices in psychology. *American Psychologist*, 1970, 25, 1094-1098.
- HAYS, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- KOONTZ, E. D. The progress of the woman worker: An unfinished story. *Issues in Industrial Society*, 1971, 2, 29-31.
- KORMAN, A. K. Toward a hypothesis of work behavior. *Journal of Applied Psychology*, 1970, 54, 31-41.
- KRAMER, C. Y. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, 1956, 12, 307-310.
- KREPS, J. *Sex in the marketplace: American women at work*. Baltimore: Johns Hopkins Press, 1971.
- MACCOBY, E. E. (Ed.) *The development of sex differences*. Stanford: Stanford University Press, 1966.
- MEGARTEE, E. I. Influence of sex roles on the manifestation of leadership. *Journal of Applied Psychology*, 1969, 53, 377-382.
- NUNNALLY, J. C. *Psychometric theory*. New York: McGraw-Hill, 1967.
- ORTH, C. D., & JACOBS, F. Women in management: Pattern for change. *Harvard Business Review*, 1971, 49, 139-147.
- ROSENKRANTZ, P., VOGEL, S., BEE, H., BROVERMAN, I., & BROVERMAN, D. M. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 1968, 32, 287-295.
- SCHEIN, V. E. The woman industrial psychologist: Illusion or reality? *American Psychologist*, 1971, 26, 708-712.
- Women in the work force. *Management Review*, 1970, 59, 20-23.
- WYLIE, R. *The self concept*. Lincoln: University of Nebraska Press, 1961.

(Received September 20, 1971)

## ETHNIC GROUP DIFFERENCES IN RELATIONSHIPS AMONG CRITERIA OF JOB PERFORMANCE<sup>1</sup>

ALAN R. BASS AND JOHN N. TURNER<sup>2</sup>

*Wayne State University*

A study was conducted to investigate racial discrimination and differential bias in criterion measures for black and white tellers in a large bank. Six supervisory ratings, and four objective criteria were obtained. Results indicated that mean differences between black and white employees on the criterion measures were generally small, and most differences that were statistically significant were reduced to nonsignificance when the effects of age and job tenure were removed. However, further analyses showed that the white supervisors based their evaluations of subordinates on objective data for black employees considerably more than they did for white employees. The results were discussed in terms of implications for criterion measurement and personnel selection.

Relatively little research dealing with the problem of discrimination in employment testing has been concerned with the problem of possible unfair discrimination with respect to criteria of job performance. A series of studies conducted by the Educational Testing Service (Flaugh & Norris, 1969; Rock & Evans, 1969) found that mean supervisory ratings as well as relationships between supervisory ratings and a job knowledge test criterion were, in part, a function of the particular rater-ratee ethnic group combination. Kirkpatrick, Ewen, Barrett and Katzell (1968) reported one study in which they suggested that supervisory ratings obtained for research purposes may have been less biased (with respect to differences between racial subgroups) than a rating scale that was used by the company as a basis for salary actions. Another study (Wollowick, Greenwood, & McNamara, 1969) obtained somewhat similar results, finding greater differences between black and white employees for a salary criterion than for a supervisory ranking criterion obtained for research purposes.

<sup>1</sup> We wish to thank the vice president for personnel, the employment manager, and the vice president for branch operations of the participating bank for their cooperation in conducting this study. We also wish to thank Ross Stagner and Thomas Hollmann for helpful comments on the manuscript.

Requests for reprints should be sent to Alan R. Bass, Department of Psychology, Wayne State University, Detroit, Michigan 48202.

<sup>2</sup> Now at the Ford Motor Company.

It seems clear, as Einhorn and Bass (1971), for example, have pointed out, that "... even if tests are used appropriately, ... discrimination in selection decisions can still occur if the criterion measures themselves are 'biased' or 'unfair' with respect to different subgroups [p. 262]." It is also clear, then, that research is necessary to investigate the extent to which criterion measures do discriminate unfairly with respect to different ethnic or racial groups. As Anastasi (1968) has pointed out, the mere presence of statistically significant differences between members of different subgroups on some measure does not, in itself, indicate unfair discrimination with respect to that measure. Thus, the mere fact that black employees might obtain lower scores, on the average, than whites on a supervisory rating criterion does not necessarily prove that the criterion measure is unfairly discriminatory with respect to black employees. Only if these differences on the criterion measure are not associated with "true" differences in job performance could the criterion measure be said to be biased or unfairly discriminatory.

The major objective of the present study was to determine the extent to which ratings of black and white employees by white supervisors are biased or unfairly discriminatory against blacks. In order to do this, it is necessary to examine not merely mean differences between black and white employees with respect to these supervisory ratings but also

the extent to which such differences are "valid" differences with respect to actual job performance.

Problems involved in establishing the validity of supervisory ratings are well known. It has been found that such ratings generally have little relationship with more objective indexes of job performance (cf. Hausmann & Strupp, 1955; Seashore, Indik, & Georgopoulos, 1960). It is possible, however, that the "validities" of supervisory ratings for predicting objective job performance measures may be different for black and white employees. The present study afforded a unique opportunity to test this possibility, since supervisory ratings as well as more objective job performance indexes were available for a large group of bank tellers. The present study, therefore, is primarily concerned with the following questions:

1. To what extent do black and white employees differ on supervisory ratings as well as on more objective performance measures?
2. To what extent do the supervisory ratings represent a biased or unfairly discriminatory criterion measure with respect to race of employee?

## METHOD

### Subjects

Subjects for this study were 244 part-time tellers (32 black and 212 white) and 190 full-time tellers (43 black and 147 white), employed by a large bank with numerous branches throughout a large metropolitan area. The part-time tellers worked an average of three days per week. All of the tellers performed the same basic job, which primarily involved face-to-face transactions with customers at the teller's window and balancing of each day's transactions at the end of the day. In addition, 163 supervisors of these tellers supplied performance evaluations for this study.

### Measures

Ratings on five separate job performance factors, and on an overall effectiveness scale, were obtained for each teller. The rating scales were constructed for use in this study on the basis of an extensive job analysis of the tellers' duties and responsibilities. The five job factors rated were (a) customer relations, which was concerned with the teller's efforts to satisfy his customer's requirements; (b) ability to sell new accounts and services; (c) quality of work, which was defined as the extent to which the teller was accurate in balancing each day's transactions; (d) alertness to irregularities, defined as the employee's alertness in detecting bad checks,

forgeries, etc.; and (e) cooperation with others, defined as the employee's effectiveness in working with his fellow employees.

Each scale was preceded by a short paragraph defining the job performance factor to be rated, followed by a 5-point continuum, with the points anchored as follows: outstanding, more than satisfactory, satisfactory, less than satisfactory, and unacceptable. A sixth, overall effectiveness rating scale was also obtained for each teller, with the same anchors as above but distributed over a 10-point scale.

The tellers were rated by their two immediate superiors, usually the branch manager and the assistant branch manager, or by the assistant manager only, when the manager was not sufficiently well acquainted with all of the tellers to be able to rate their performance. Of the 434 tellers in the study, 368 or 84.9% were rated by two supervisors, and the remaining 66 tellers were each rated by only one supervisor. Where two supervisory ratings were obtained, the two ratings on each scale were averaged for each teller. Generally, there was good agreement between the two raters, with the interrater reliabilities (corrected by Spearman-Brown) ranging from .69 to .83 for the six rating scales, with a median reliability of .76.

In addition to the ratings, four other criterion measures were available for each teller. Two of these were error counts—number of shortages and number of overages—obtained over a one-month period for each teller.<sup>3</sup>

The third non-rating criterion measure was an attendance figure computed for each full-time teller. This was obtained by dividing the number of days an employee potentially could have worked during the calendar year preceding the data collection for this study by the number of days he actually did work, yielding a percent-of-time worked index. This index was not computed for the part-time tellers.

The fourth nonrating criterion measure was an adjusted salary increase index. This index was computed by obtaining the difference between the teller's present salary and the salary that the teller was predicted to have on the basis of his starting salary and his length of service with the bank. Thus, this difference can be interpreted as reflecting the net salary increase that the teller had received since entering his job beyond that which would be accounted for solely on the basis of his starting salary

<sup>3</sup> Each of these errors is detectable from the daily balance sheet that the teller prepares, indicating either that the teller entered an amount in a transaction incorrectly or that the teller made an error in counting out money for a customer. An overage occurs when the teller's balance at the end of the day is greater than it should be, while a shortage indicates a balance less than should be the case. If the error involves a money transaction rather than a transcribing or arithmetic error, an overage implies that some customer received less than should have been the case in a transaction, while a shortage suggests that a customer received more money than he should have received.

TABLE 1  
INTERCORRELATIONS AMONG CRITERION MEASURES AND CONTROL VARIABLES FOR FULL-TIME TELLERS

Variable	1	2	3	4	5	6	7	8	9	10	11	12	Race <sup>a</sup>	Race <sup>b</sup> (Partials)
Supervisor's ratings														
Customer relations (1)	—	56	35	59	70	79	14	22	-13	-20	10	20	19*	.07
New accounts (2)	60	—	31	31	40	42	12	12	-18	-25	18	14	.04	-.07
Quality of work (3)	33	29	—	49	38	61	37*	17	-55**	-51**	33*	03	.06	.00
Alertness (4)	54	54	68	—	35	84	16	22	-17	-23	35*	02	-.01	-.13
Cooperation (5)	59	34	49	55	—	62	12	14	-04	-09	04	21	.14	.15
Overall effectiveness (6)	68	58	74	78	75	—	28	33	-18	-21	25	12	.12	.04
Adjusted salary increase (7)	28**	15	29**	30**	26**	36**	—	51**	-20	-24	27	11	.11	.08
% of time worked (8)	12	.09	15	20*	.03	15	12	—	-06	-12	36*	.02	.28**	.17*
Number of shortages (9)	-18	-10	-45**	-32**	-10	-32**	-17	-27**	—	75**	-19	.08	-32**	-24**
Number of overages (10)	.07	.01	-08	.06	.09	.02	10	-05	21	—	-21	.15	-18*	-.14
Time on job (11)	14	21*	13	25**	-06	11	-10	35**	-13	-06	—	-05	.32**	
Age (12)	36**	26**	.08	25**	10	15	15	17*	.07	.03	25	—	.30**	

Notes. Decimal points are omitted. Correlations for black tellers above diagonal, for white tellers below diagonal. N for black tellers ranged from 29 to 38, for white tellers from 112 to 129 due to missing data.

<sup>a</sup> Scored by assigning black = 0, white = 1.

<sup>b</sup> Partial correlations between race and criterion measures with age and time on job partialled out.

\*  $p < .05$ .

\*\*  $p < .01$ .

TABLE 2  
INTERCORRELATIONS AMONG CRITERION MEASURES AND CONTROL VARIABLES FOR PART-TIME TELLERS

Variable	1	2	3	4	5	6	7	8	9	10	11	Race <sup>a</sup>	Race <sup>b</sup> (Partials)
Supervisor's Ratings													
Customer relations (1)	—	.59	.34	.23	.39	.50	.14	-.11	-.08	.12	.29	.13*	.07
New accounts (2)	.59	—	.11	.04	.27	.29	.23	.29	.00	.33	-.18	.09	.05
Quality of work (3)	.48	.47	—	.56	.55	.87	.28	-.62**	-.48**	.32	.07	.24**	.21**
Alertness (4)	.55	.53	.73	—	.51	.64	.54**	-.30	-.27	.32	-.05	.16*	.10
Cooperation (5)	.65	.43	.48	.51	—	.72	.32	-.11	.06	.15	-.17	.13*	.12
Overall effectiveness (6)	.76	.66	.70	.77	.75	—	.35	-.46**	-.36*	.35	-.11	.22**	.17*
Adjusted salary increase (7)	.16*	.11	.26**	.26**	.12	-.21*	—	-.04	-.12	.75	-.14	.18*	.17*
Number of shortages (8)	.07	-.04	-.22**	-.09	.06	-.10	-.10	—	.62	-.09	-.16	-.18*	-.19*
Number of overages (9)	-.07	-.06	-.11	-.05	-.09	-.11	.01	.52**	—	-.25	-.08	-.06	-.09
Time on job (10)	.20**	.17*	.18*	.28**	.14*	.23**	-.07	.09	.07	—	-.22	.20**	
Age (11)	.00	.03	-.04	.02	.03	.01	.03	.03	-.07	.02	—	.27**	

Note. Decimal points are omitted. Correlations for black tellers above diagonal, for white tellers below diagonal. N for black tellers = 26; for white tellers from 193 to 201 due to missing data.

\* Scored by assigning black = 0, white = 1.

<sup>b</sup> Partial correlations between race and criterion measures with age and time on job partialled out.

\*  $p < .05$ .

\*\*  $p < .01$ .

and the length of time he had been on the job. An index comparable to this has been suggested by Hulin (1963).

Besides the ten criterion measures described above (six supervisory ratings and four objective scores) three additional control variables were obtained for each employee. These were tenure—computed as the length of time, in weeks, that the employee had worked for the bank, age, and race.

## RESULTS

Correlations among the six rating scales, the four objective criterion measures, and the three control variables computed separately for black and white full-time and part-time employees, are presented in Tables 1 & 2. As would be expected, the data show moderate to high correlations among the rating scales, for both part-time and full-time employees, indicating lack of independence among the performance traits and/or halo error. Also, for both part-time and full-time employees there are generally low and nonsignificant correlations between the ratings and the objective criterion measures. One notable exception is the fairly high degree of relationship between ratings of "quality of work" and the number of shortages and overages. This provides some evidence of validity for these ratings, since "quality of work" was defined for the raters as the employee's accuracy in balancing his accounts. Employee tenure is related to performance ratings, especially for part-time employees, with longer-tenure employees generally obtaining higher job performance ratings. Age is related to the job performance measure for the white full-time employees but not for the other groups.

With respect to the question of racial differences on the criterion measures, it will be noted that race was significantly related to both performance ratings and objective measures for the part-time employees, and primarily to the objective measures for the full-time employees. In almost every case white employees exhibited higher average scores than did the black employees, although the magnitude of these racial differences is relatively small (the correlations—although statistically significant—are relatively small, and the mean differences are also small, generally no more than half a scale point for the performance ratings). The means and standard

deviations of these measures for black and white employees are presented in Table 3.

Even though whites appear to obtain higher mean scores than blacks on these criterion measures, it is possible that the differences are due, in part, to differences in age and job tenure between the two groups. As Table 3 indicates, whites are significantly older and have more job experience than do the blacks in this sample, and age and job tenure are both correlated with the various criterion measures to some extent. Therefore, partial correlations were obtained between race and the criterion measures, in which both age and tenure were partialled out. The results of these analyses are presented in the last column of Tables 1 and 2. For full-time tellers, race was significantly correlated with four of the criterion measures before partialling, and with two of them (percent of time worked and number of shortages) after both age and tenure were partialled out. For part-time tellers, race was significantly correlated originally with seven of the criterion measures; four of these relationships were still significant after age and tenure were partialled out (quality of work, overall effectiveness, number of shortages, and adjusted salary increase). Thus, while removing the effects of age and job tenure does tend to reduce the differences between job criterion measures for black and white employees, some significant differences between these two racial groups still remain beyond those which can be accounted for by age and tenure, although these significant differences are generally quite small.

Even if black employees consistently score lower on these job performance measures than do whites, it is necessary to determine the extent to which the mean differences in the supervisor's performance ratings may be reflecting unfair bias or discrimination against black employees, especially since virtually all of the supervisors in the study were white. In order to investigate this question, correlations between the supervisory ratings and the objective criterion measures were examined separately for black and white employees, both full-time and part-time. These correlations are included in Tables 1 and 2. Here it can be seen that there is generally a larger

TABLE 3

MEANS AND STANDARD DEVIATIONS OF CRITERION AND CONTROL VARIABLES

Variable	Group <sup>a</sup>	Full-time tellers			Part-time tellers		
		M	SD	n	M	SD	n
Customer relations	B	3.33	.84	38	3.33	.66	26
	W	3.67*	.72	128	3.60*	.69	207
New accounts	B	3.30	.70	38	3.04	.73	26
	W	3.36	.75	128	3.24	.71	207
Quality of work	B	3.30	.92	38	2.96	.86	26
	W	3.41	.80	128	3.55**	.72	207
Alertness to irregularities	B	3.49	.78	38	3.08	.54	26
	W	3.46	.74	128	3.39*	.59	207
Cooperation	B	3.54	.85	38	3.37	.69	26
	W	3.81	.82	128	3.69*	.78	207
Overall effectiveness	B	6.51	1.73	38	5.79	1.46	26
	W	6.97	1.56	128	6.82**	1.44	207
Percentage of time worked	B	97.75	2.22	36	—	—	—
	W	98.80**	1.19	145	—	—	—
Number of shortages	B	6.03	4.51	40	3.75	3.81	32
	W	3.56**	2.59	130	2.69**	2.56	199
Number of overages	B	3.93	2.80	40	3.07	2.98	32
	W	2.40*	3.82	130	1.80	6.70	200
Adjusted salary increase	B	98.76	5.64	39	97.19	5.04	32
	W	100.46	6.24	128	100.30*	5.52	200
Time on job	B	109.70	82.90	43	69.41	39.31	27
	W	194.10**	113.70	147	156.90**	142.96	210
Age	B	23.70	5.60	43	27.48	8.39	27
	W	29.80**	9.90	147	34.87**	8.53	212

<sup>a</sup> B = black tellers; W = white tellers.\* Mean differences for black and white tellers significant at  $p < .05$ .\*\* Mean difference for black and white tellers significant at  $p < .01$ .

relationship between the performance ratings and the objective measures (particularly overages and shortages) for black than for white employees, especially for the part-time tellers. Of particular interest here are the correlations between ratings on "quality of work" and number of shortages and overages. It would be expected that ratings on work quality should correlate with overages and shortages, since this rating scale was defined as the extent to which the teller was accurate in balancing each day's transactions. Four comparisons between correlations for black and white tellers are relevant here—viz., correlations between quality of work ratings and shortages and overages for both part-time and full-time tellers. In all four of these comparisons, the correlation between the ratings and number of errors is higher for blacks than for whites, and for three of the four comparisons (viz., shortages and overages for part-time tellers and overages for full-time tellers) the differences between correlations

for blacks and whites were statistically significant ( $p < .05$ ). Similarly, the correlations between the "overall effectiveness" ratings and shortages and overages are generally higher for blacks than for whites, although these differences do not attain statistical significance. (For part-time tellers, the difference in correlations between overall effectiveness ratings and number of shortages for black and white employees approached significance— $p < .10$ .) The direction of the differences in correlations is similar for percent of time worked and for salary increases.

Finally, it is of interest to note the correlations between the adjusted salary increase measure and the other criterion measures. For full-time employees, the correlation between salary increases and attendance record is significantly higher for blacks than for whites ( $p < .05$ ). Unfortunately, attendance data were not available for part-time employees so that we can not examine the comparable relationship for that group. Further, salary

increases were found to be significantly related to a number of the supervisory rating scales (i.e., nonobjective criterion measures) for white employees, but to only one rating scale for blacks, for both part-time and full-time employees.

Thus, in summary, the findings suggest that correlations between supervisory ratings and objective criterion measures tend to be higher for black than for white employees, and also that the correlation between salary increases and an objective criterion (attendance) is higher for black than for white employees.

#### DISCUSSION

To what extent are the supervisory evaluations obtained here biased against black employees? Considering just the performance ratings, the data indicate that there are no significant mean differences between black and white full-time tellers when age and job tenure are held constant, and only two ratings (Quality of Work and Overall Effectiveness) significantly differentiate black and white part-time tellers with age and job tenure held constant. Even here, moreover, the mean differences are quite small and of little practical significance. With regard to salary increases, which may be considered to be a more general or more "ultimate" supervisory evaluation, there was no significant mean difference for full-time tellers and again a very slight mean difference (in favor of the white employees) for part-time tellers. Thus, in terms of mean differences between black and white employees our data do not indicate any direct, systematic bias against black employees with respect to supervisory evaluations of these employees.

At the same time, it seems reasonable to conclude that some differential criterion bias does occur here. For example, the results for the full-time tellers indicate that salary increases were based on attendance records for blacks but not for whites. In addition, salary increases tended to be significantly correlated with supervisory ratings for white tellers but the corresponding correlations for black tellers were generally non-significant. Thus, these data suggest that while the supervisors tended to rate the performance of black and white employees quite similarly, on the average, they had a tendency to consider

more objective aspects of performance when making salary recommendations for blacks but to consider other, less objective, factors when making salary recommendations for whites.

The results for the part-time tellers are also generally consistent with the interpretation that objective performance measures are considered more important in evaluating black workers. Here the correlations between salary increase and other performance measures are about equal for blacks and whites, but the ratings themselves are more strongly related to errors (the only objective measures available for part-time tellers) for blacks than for whites.

In evaluating these findings, the possibility was considered that the lower relationships between supervisory evaluations and objective data for whites might be due to the somewhat restricted range and skewed distributions of objective scores for whites. There were three correlations where such restriction and skewness might have been a factor: salary increases versus attendance and "quality of work" ratings versus overages for white full-time tellers; and "quality of work" ratings versus overages for white part-time tellers. Using the procedure suggested by Carroll (1961), it was found that the maximum correlations attainable, given the frequency distributions in these cases, were .93, .94, and .86, respectively. Thus, it seems clear that these supervisors simply did not base their evaluations of white employees on objective data even though they could have done so, while they did base their evaluations of black employees on objective data.

One possible explanation for these findings is simply that white supervisors are quite sensitive to the existence of racial tension and are concerned about the possibility of being accused of biasing their evaluations against blacks, and thus rely heavily on those aspects of performance that have been recorded and of which they can be relatively certain in evaluating black workers. Several studies have obtained results that are relevant to our findings and tend to lend some support to this explanation as well as to the generality of our results.

In a study mentioned earlier, Flaugh and

Norris (1969) found that ratings of job knowledge given by white supervisors were more highly correlated with actual job knowledge test scores for black than for white subordinates ( $r = .46$  and  $.30$ , respectively). While this was an incidental finding in their study and no test of significance was reported, the direction of the difference is certainly consistent with our data, and is especially comparable if job knowledge can be considered to be reflected in objective or overtly observable aspects of performance.

In a laboratory study, Rotter and Rotter (1966) found that white raters generally gave higher ratings to black than to white workers when performance was poor, but found no differences in ratings given to black and white workers when performance was good. They interpret their results as suggesting that evaluators experience guilt over a low rating given to a minority group member, so that to avoid guilt feelings the evaluator leans over backwards to be fair or lenient when rating poor performance of minority group members. If this tendency to be lenient is due to a reluctance to rate minority group members low in the absence of definitive information, at the risk of appearing racially biased, then the Rotter and Rotter findings and interpretations are clearly applicable to the present data as well.

In interpreting the results of a study by Dienstbier (1970) with white male high school students, and an unpublished laboratory study by Pass (1971) using college white male undergraduates, Leventhal (1971) has suggested that white evaluators (especially pro-Negro whites) have a tendency to display either highly positive or highly negative reactions to black workers, with the direction of the reaction depending on whether an individual black is behaving in a socially approved fashion. If we assume that good attendance records and low error rates represent socially approved behaviors, then these laboratory findings cited by Leventhal also appear to be consistent with the present field study findings.

Another possible explanation for our findings is that white supervisors feel more psychologically and/or socially distant from black employees than from white employees and therefore non-objective factors such as

interpersonal attraction simply have less opportunity to influence the evaluations. A good deal of research has indicated that supervisory ratings are considerably influenced by interpersonal considerations such as supporting behaviors, ingratiation attempts, interpersonal attraction, similarity of attitudes and values, etc. (see, e.g., Hausmann & Strupp, 1955; Kallejian, Brown & Weschler, 1953; Kipnis, 1960; Miles, 1964). We do not know from our data what factors other than objective performance were used by the present supervisors in evaluating white employees. As an incidental variable, however, supervisors were simply asked to indicate which of two factors they considered most important in evaluating subordinates: (a) production records, e.g., shortages and overages or (b) attitude and motivation. The great majority of the supervisors (81 or 70%) indicated that the subordinates' attitudes and motivation were most important, although it seems clear from our data that this may have been true for white subordinates but not for blacks.

Can we conclude that the performance evaluations obtained here were more fair for black than for white employees? We think not. While it is true that the ratings given to blacks are more closely related to objective data, it is also true that they may fail to take into account possible compensatory factors (e.g., motivation, effort, attitude, interpersonal factors, etc.) for blacks, while they presumably do take these considerations into account for white employees. In their attempt to lean over backwards to be objective in their evaluation of blacks, supervisors may at the same time be failing to give black employees the "benefit of the doubt." Thus, if a white employee makes numerous errors, but at the same time exhibits other desirable job behaviors, he would apparently receive a higher evaluation (and probably even higher merit increases) from his supervisor than would the black employee with similar characteristics.

Finally, what are the implications of these results for test validation and the fair use of selection tests? If supervisory ratings were used as criterion measures in this situation, and differential validity was investigated, it seems likely that different kinds of tests would predict "success" for black and white

employees. Aptitude tests would most likely be found to be related to ratings of black employees and personality or attitude measures might be found to be predictive of success for white employees, due to differences in the nature of the ratings for the two groups. It would clearly be inappropriate, however, to use different criterion measures for evaluation of black and white employees doing the same job. Therefore, in the present situation, the supervisory ratings, as obtained, would clearly be inappropriate as criterion measures against which to validate tests, even if the validation were done separately for the two groups. Rather, it is necessary for the organization to define clearly what constitutes success on the job, and then to make sure that criterion measures used to assess job performance are equivalent for all employees. To insure equivalence, one could concentrate only on those elements where equivalence across groups is known, viz., objective data. However, objective data may be difficult to obtain and may also be deficient as criteria, since there may be more to most jobs than just the number of units produced, number of errors recorded, or the attendance record. A more attractive alternative, then, would be to attempt to increase the equivalence of the ratings either by refining the rating scales (cf. Smith & Kendall, 1963), by using multiple raters, and/or by training the raters in the nature and use of the rating scales (cf. Brown, 1968). Then, ratings as well as objective data could be used as multiple criterion measures in conducting test validation studies that would be fair to both groups.

It is possible that in studies that have found tests to be differentially valid for black and white employees, the reasons for the differential validity lie in differences in the nature and meaning of the criterion measure used rather than differences in the "meaning" of the test scores. Future research concerned with differential validity of selection tests should attempt to investigate this possibility.

#### REFERENCES

- ANASTASI, A. *Psychological testing*. New York: Macmillan, 1968.
- BROWN, EVA M. Influence of training, method, and relationship on the halo effect. *Journal of Applied Psychology*, 1968, 52, 195-199.
- CARROLL, J. B. The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 1961, 26, 347-372.
- DIENSTBIER, R. A. Positive and negative prejudice: Interactions of prejudice with race and social desirability. *Journal of Personality*, 1970, 38, 198-215.
- EINHORN, H. J., & BASS, A. R. Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, 1971, 75, 261-269.
- FLAUGHER, R. L., & NORRIS, L. Ethnic group membership as a moderator of supervisory ratings. Paper presented at the Convention of the American Psychological Association, Washington, D.C., September 1969.
- HAUSMANN, H. J., & STRUPP, H. H. Non-technical factors in supervisors' ratings of job performance. *Personnel Psychology*, 1955, 8, 201-217.
- HULIN, C. J. Relevance and equivalence in criterion measures of executive success. *Journal of Industrial Psychology*, 1963, 1, 67-78.
- KALLEJIAN, V., BROWN, P., & WESCHLER, I. R. The impact of interpersonal relations on ratings of performance. *Public Personnel Review*, 1958, 14, 166-170.
- KIRKPATRICK, J. J., EWEN, R. B., BARRETT, R., & KATZELL, R. A. *Testing and fair employment*. New York: New York University Press, 1968.
- KIPNIS, D. Some determinants of supervisory esteem. *Personnel Psychology*, 1960, 13, 377-391.
- LEVENTHAL, G. S. Equity and reward allocation in social relationships. Research proposal submitted to the National Science Foundation, 1971.
- MILES, R. E. Attitudes toward management theory as a factor in managers' relationships with their superiors. *Academy of Management Journal*, 1964, 7, 308-314.
- PASS, J. Anti-Negro and pro-Negro prejudice and the perception of equity. Unpublished master's thesis, North Carolina State University, 1971.
- ROCK, D. A., & EVANS, F. R. Aptitude and rating factors of Negroes and whites. Paper presented at the Convention of the American Psychological Association, Washington, D.C., September, 1969.
- ROTTER, N., & ROTTER, G. S. Race, work performance, and merit rating: An experimental evaluation. Paper presented at the Convention of the Eastern Psychological Association, Philadelphia, April 1969.
- SEASHORE, S. E., INDIK, B. P., & GEORGOPoulos, B. S. Relationships among criteria of job performance. *Journal of Applied Psychology*, 1960, 44, 195-202.
- SMITH, P. C., & KENDALL, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 1963, 47, 149-155.
- WOLLOWICK, H. B., GREENWOOD, J. M., & McNAMARA, W. J. Psychological testing with a minority group population. *APA Proceedings*, 77th Annual Convention, 1969.

# PREDICTING THE EFFECTS OF LEADERSHIP TRAINING AND EXPERIENCE FROM THE CONTINGENCY MODEL: A CLARIFICATION<sup>1</sup>

FRED E. FIEDLER<sup>2</sup>

*University of Washington*

An article published in the April 1972 issue of the *Journal of Applied Psychology* presented a new interpretation of leadership training and experience, as well as supporting empirical findings. This article attempts to clarify a number of points that the article did not make sufficiently clear and, therefore, correct the various misinterpretations of the findings as well as of the underlying theory.

The future inventor of a method for "de-publishing" disowned statements in journal articles will surely be acclaimed as one of mankind's great benefactors. Several anonymous comments forwarded to me by the editor make it quite obvious that several parts of my recent article (Fiedler, 1972a) were unclear and could well benefit from such depublication procedures. In lieu of this more elegant but as yet unavailable solution, I am most grateful to the editor for giving me the opportunity to rectify this problem.

My article reported that leadership training and experience had opposite effects on the performance of relationship- and task-motivated leaders, with the direction of the effect depending on the favorableness of the leadership situation. The research was based on the Contingency Model (Fiedler, 1964, 1967, 1971) which predicts that task-motivated leaders (low LPC) perform best in very favorable and in unfavorable situations while relationship-motivated leaders (high LPC) perform best in moderately favorable situations. Situational favorableness has been defined as the degree to which the situation gives the leader control and influence, assuming, however, ". . . that the leader and the group members have the required physical resources, skills, and abilities . . ." (Fiedler, 1967, p. 22). In other words, the classification is based on the assumption that the leader has already acquired the required skills

and abilities, whether by training or by experience.

The hypothesis of my article, restated in different form, was that training and experience typically assist the leader to develop better relations with his group and a better working knowledge of the job. Hence, training and experience will tend to increase the leader's control and influence, and thus, by definition, the situational favorableness. This hypothesis implies, corollarily, that a relatively inexperienced and untrained leader will perform *as if the situation were less favorable for him than for the highly trained and experienced leader*. It follows that a situation that is favorable for the trained and experienced leader is likely, therefore, to be only moderately favorable for the inexperienced leader; a situation that is moderately favorable for the trained and experienced leader will be unfavorable for the inexperienced and untrained leader; and a situation that is unfavorable for the trained and experienced leader will be highly unfavorable for the untrained and inexperienced leader. These hypotheses are summarized in Table 1.

A preliminary study, using data from a number of earlier investigations, yielded the median correlations shown on Table 2. (For the complete table, identifying subsamples and indicating N's, see Fiedler, 1972a.) These summarize the relationship between group performance and the number of years of experience or training received by the group's leader with high and low LPC scores, and with training intended for situations which would be classified as favorable, moderate, or unfavorable for trained and experienced lead-

<sup>1</sup> This article is published out of turn at the request of the editor.

<sup>2</sup> Requests for reprints should be sent to Fred E. Fiedler, Organizational Research, University of Washington, Seattle, Washington 98195.

TABLE 1  
SUMMARY OF HYPOTHESES REGARDING THE EFFECTS OF TRAINING AND EXPERIENCE

Favorableness of situation for trained and experienced leader	Performance level of leaders with adequate training and experience		Favorableness of situation for inadequately experienced leader	Performance level of leaders without adequate training and experience		Predicted effect of training and experience for previously untrained leader		
	LPC			LPC				
	High	Low		High	Low			
Very favorable	Poor	Good	Moderate	Good	Poor	High decreases Low increases		
Moderately favorable	Good	Poor	Unfavorable	Poor	Good	High increases Low decreases		
Unfavorable	Poor	Good	Very unfavorable	Good?	Poor?	High decreases Low increases		

ers. Four new studies (Csoka & Fiedler, 1972) provide further support for these hypotheses. At issue in this note is the interpretation of these findings.

As mentioned in my previous article, "The original classification of situational favorableness assumed technically qualified leaders. The leader who is inexperienced or untrained would, of course, find the same situation less favorable (Fiedler, 1972a, p. 115)." It is apparent that this was not sufficiently emphasized since a number of readers seemingly missed this important point. This also made Figure 1 confusing because the arrows were intended to show what will happen to the *untrained* leader as a result of training.

Rather than trying to explain the ill-fated Figure 1 of my earlier article, a new, and hopefully improved, Figure 1 illustrates the same point with actual data obtained from companies of a federation of consumer cooperatives (Fiedler, 1967, 89-107). The subjects were the general managers of the various companies. Some of them had extensive experience and training in several other companies in which they might have served as assistant or sales managers prior to being selected for the general management of their present company. The leadership situation had previously been rated as having relatively high task-structure and position power (Fiedler, 1967, p. 134). The criterion here used was the "percent of operating efficiency" (essen-

tially overhead costs) computed as a ratio of total sales. Figure 1 shows the average performance scores of high LPC and low LPC managers with relatively much and relatively little experience.

As mentioned earlier, the situation was rated as relatively favorable for the experienced managers. Just as the model predicts, the experienced managers with low LPC performed better than those with high LPC (right-hand side of the graph). We would further predict that the situation would be only *moderately favorable for the inexperienced managers*, and that the high LPC managers would, therefore, perform better than the low LPC managers, as was the case. We now infer that the low LPC managers' per-

TABLE 2  
MEDIAN CORRELATIONS BETWEEN AMOUNT OF EXPERIENCE OR TRAINING AND GROUP PERFORMANCE

Situation	Least preferred co-worker score	
	High	Low
Very favorable for the trained leader	-.21*	.28**
Moderately favorable for the trained leader	.84**	-.14
Unfavorable for the trained leader	-.74**	.60

*Note.* See Fiedler, 1972a for nature and size of subsamples.

\*  $p < .05$ .

\*\*  $p < .01$ .

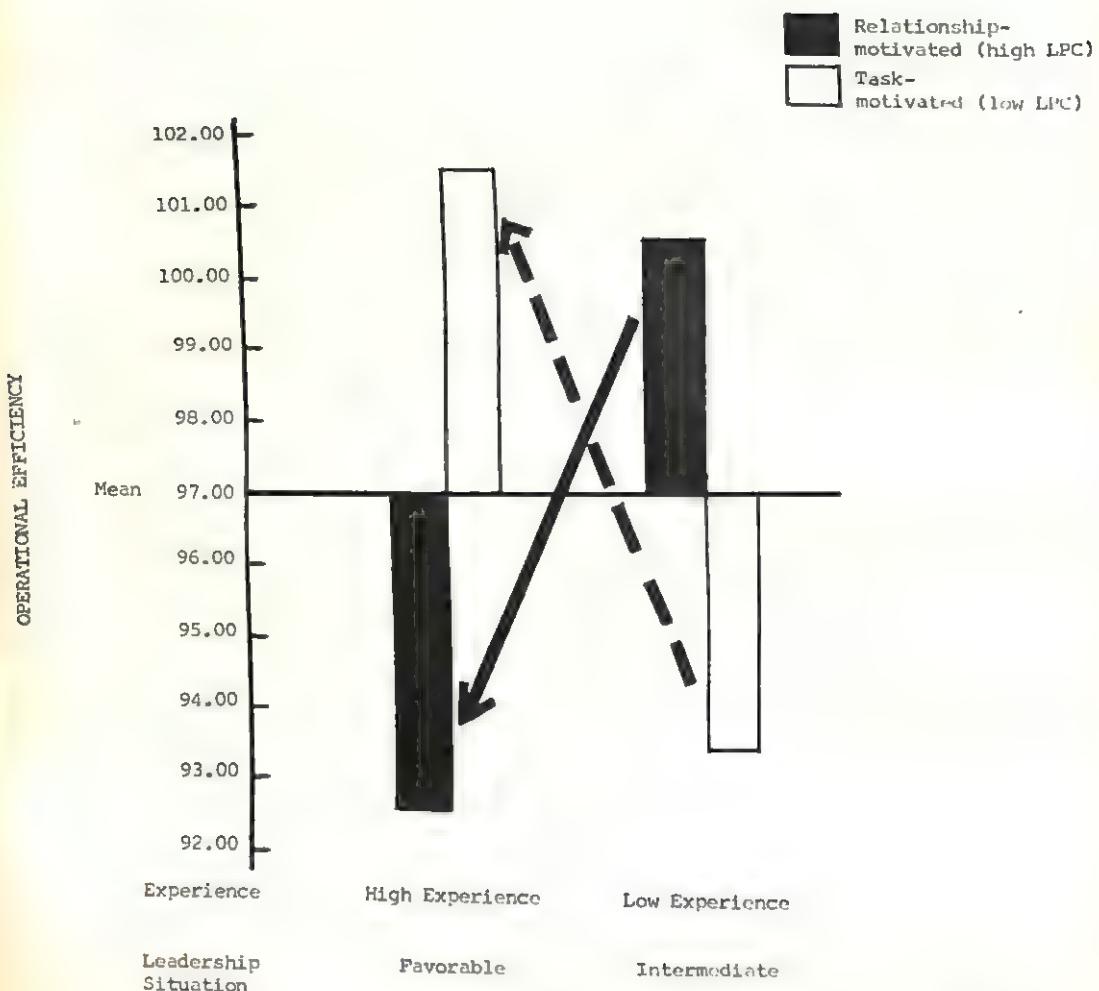


FIG. 1. Mean performance scores (percent operating efficiency/sales volume) of relationship- and task-motivated general managers with relatively high and low levels of experience in the cooperative federation.

formance increased as they gained in experience over the years, and that the high LPC managers' performance decreased during that time. This is indicated by the arrows. It is worth noting that the high LPC leaders with relatively little experience actually performed better than the high LPC managers with relatively much experience, and they also performed as well as did the low LPC managers with experience.

An alternative hypothesis would be that the increased experience and training changed the leaders' LPC scores rather than changing the actual or perceived situational favorability. However, this explanation does not seem as plausible in light of other data. For example, a study of school principals by Mc-

Namara (1968) showed that the leaders' relations with staff members improved over time and so did his staff members' rating of his competence. However, the LPC scores of principals remained fairly stable over an 18 months period. It seems, therefore, less likely that LPC changed than that the favorableness changed. This is also illustrated by Chemers' (1969) study which showed that a four-hour human relations program preparing leaders who work in culturally mixed groups, differentially changed the behavior of high and of low LPC leaders while a control program resulted in no differences in leader behavior. Although the possibility cannot be ruled out, it seems unlikely that a four-hour

training program would have lasting effects on a person's motivational system.

As some readers have remarked, the model does assume situations even less favorable than Octant VIII (poor leader-member relations, low task structure, weak position power). That this is the case has already been shown in several of our previous studies (e.g., Fiedler, 1966; Meuwese & Fiedler, 1964). There has been a suggestion in some of these data that extremely unfavorable situations might simply wash out the effects of different leadership styles or motivations or that a more relationship-motivated leadership might be called for (Fiedler, 1967, p. 207). As yet, the evidence is clearly insufficient to speak to this point with any degree of confidence.

A number of the Contingency Model's critics have charged that ". . . the theory keeps changing to fit the data" and that it is becoming increasingly complex. Both of these observations are accurate. But, as applied to the problem of incorporating training and experience in the situational favorableness dimension, it is perhaps worth noting that I said in 1967 that ". . . there are many other dimensions which should influence the favorableness of the situation for the leader. Thus . . . expertise of the leader and his familiarity with the task and with his group, should affect the degree to which he can influence the members of his group (Fiedler, 1967, p. 151)." The theory will, of course, continue to change as new data become available, and, in all probability, empirical research in the leadership area will continue to uncover complex interactions (e.g., Graen, et al., 1972; House, et al., 1971; Yukl, 1971). We simply have to live with the fact that any attempt to predict pretzel-shaped relationships will require the development of pretzel-shaped hypotheses.

Whether or not the theory changes or becomes more complex is, of course, quite irrelevant in the final analysis as long as it helps us to understand and to predict better the complexities of leadership. It is, therefore, particularly noteworthy that the hypothesis of the 1972 article has already been supported in four studies of 221 different military groups (Csoka & Fiedler, 1973; Fiedler,

1972b). And a recently completed laboratory experiment at the University of Utah by Martin Chemers (personal communication, December 1972) further substantiated the theory. His experiment showed under highly controlled conditions that training designed for a moderately favorable situation decreased the performance of task-motivated leaders but increased the performance of relationship-motivated leaders just as the Contingency Model predicts.

#### REFERENCES

- CHEMERS, M. M. Cross-cultural training as a means for improving situational favorableness. *Human Relations*, 1969, 22, 531-546.
- COKA, L. S., & FIEDLER, F. E. The effect of military leadership training: A test of the contingency model. *Organizational Behavior and Human Performance*, 1972, 395-407.
- FIEDLER, F. E. A contingency model of leadership effectiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 1. New York: Academic Press, 1964.
- FIEDLER, F. E. The effect of leadership and cultural heterogeneity on group performance: A test of the Contingency Model. *Journal of Experimental Social Psychology*, 1966, 2, 237-264.
- FIEDLER, F. E. *A theory of leadership effectiveness*. New York: McGraw-Hill, 1967.
- FIEDLER, F. E. Validation and extension of the contingency model of leadership effectiveness: A review of empirical findings. *Psychological Bulletin*, 1971, 76, 128-148.
- FIEDLER, F. E. Predicting the effects of leadership training and experience from the contingency model. *Journal of Applied Psychology*, 1972a, 56, 114-119.
- FIEDLER, F. E. The effects of leadership training and experience: A Contingency Model interpretation. *Administrative Science Quarterly*, 1972b, 17, 453-470.
- GRAEN, G., DANSEREAU, F., JR., & MINAMI, T. Dysfunctional leadership styles. *Organizational Behavior and Human Performance*, 1972, 7, 216-236.
- HOUSE, R. J., FILLEY, A. C., & KERR, S. A path-goal theory of leader effectiveness. *Administrative Science Quarterly*, 1971, 16, 19-30.
- MICHAEL, V. D. Leadership, staff, and school effectiveness. Unpublished doctoral dissertation, University of Alberta, Edmonton, Alberta, Canada, 1968.
- MEUWESE, W., & FIEDLER, F. E. Leadership and group creativity under varying conditions of stress. Urbana, Ill.: Group Effectiveness Research Laboratory, University of Illinois, 1965.
- YUKL, G. Toward a behavioral theory of leadership. *Organizational Behavior and Human Performance*, 1971, 6, 414-440.

(Received October 24, 1972)

## PREDICTING THE EFFECTS OF LEADERSHIP TRAINING AND EXPERIENCE FROM THE CONTINGENCY MODEL: SOME REMAINING PROBLEMS<sup>1</sup>

STEVEN KERR<sup>2</sup> AND ANNE HARLAN

*Ohio State University*

Fiedler's clarifying remarks on using the Contingency Model to predict effects of leadership training and experience have resolved many of the original article's apparent inconsistencies. Some problems still remain, however, which could seriously impair the usefulness of Fiedler's recommendations. This article briefly discusses some of these problems and suggests some possible courses of action.

While Fiedler's clarifying note concerning use of the Contingency Model to predict the effects of leadership training and experience goes a long way toward resolving the inconsistencies that seemed to afflict the original article, a number of difficulties remain. These are by no means insurmountable, but need to be addressed if the Contingency Model is to be of use in the way Fiedler claims it can be. A few of these difficulties as well as possible courses of action will be discussed below.

### *To What Extent Is Situation Favorableness Affected by Training and Experience?*

As originally designed, it seems as though situation favorableness was intended to be "immune" from such variables as leader training and experience, concentrating instead on "how the organization and the task affect the leader's ability to motivate his members and to direct and coordinate their efforts [Fiedler, 1967, p. 22]." Thus one component of situation favorableness, position power, was intended to represent "the degree to which the position itself enables the leader to get his group members to comply with and accept his direction and leadership [Fiedler, 1967, p. 22]." A second component, task structure, is spoken of by Fiedler as concerning primarily the nature of the task, which is an attribute "determined by the organization [1967, p. 29]." The third component, leader-

member relations, is probably most influenced by the personal attributes of the leader, but even here Fiedler reminds us that the nature of the organization itself, and its bestowal of legitimate position on the leader, play an important role.

In fact, however, by examining the specific scales and checklists used to compute the situation favorableness score, it is easy to see that the various items differ widely in their immunity from personal leader variables. The following checklist items for Position Power, for example, appear to be largely a function of the formal organization authority and status systems and are probably not readily changeable by subjecting the leader to training or experience:

Leader can recommend punishments and rewards.  
Leader can punish or reward members on his own accord.

Leader enjoys special or official rank and status in real life which sets him apart from or above group members.

Other items on the same checklist are probably highly dependent on the degree to which the leader is perceived by the members to be trained and experienced. For example:

Leader's opinion is accorded considerable respect and attention.  
Compliments from the leader are appreciated more than compliments from other group members.  
Leader knows his own as well as members' job and could finish the work himself if necessary [Fiedler, 1967, p. 24].

The same contrast exists between such dimensions of task structure as "solution specificity" (which measures the degree to which

<sup>1</sup> This article is published out of turn at the request of the editor.

<sup>2</sup> Requests for reprints should be sent to Steven Kerr, Faculty of Management Sciences, Ohio State University, 1775 South College Road, Columbus, Ohio 43210.

there is more than one correct solution and which is unlikely to be much affected by leader training and experience) and others such as "goal clarity" (which measures the extent to which task requirements are clearly stated and known to group members and which could probably be greatly influenced by leader training and experience). As mentioned earlier, the third component of leader-member relations is also a mixed bag of elements that are susceptible to changes in leader variables and elements which are not.

If leader training and experience are to be examined by the Contingency Model for the purpose of predicting their effects on performance, and if determination of situation favorableness is an important step (as seems to be the case, see Fiedler clarification, Table 1) in such an examination, it seems that greater attention needs to be paid to the question of just how influencing the situation favorableness score is to changes in leader experience and training. Alternatively, it might be ideal to revise the existing measures of position power, task structure, and leader-member relations to make them as independent of personal leader variables as is possible. In this way situation favorableness would more closely serve its original purpose, that of determining how the *organization and the task* affect the leader's ability to motivate his members and to direct and coordinate their efforts. Amount of leader experience and training might then be examined as a moderating variable, without causing concern as to how much it is causing the situation favorableness score to be modified.

#### *Is "Leader Experience and Training" an Important Additional Variable to the Contingency Model?*

While Fiedler's contention that the effects of leader experience and training can be predicted from the Contingency Model is an interesting and challenging one, his argument that the situation favorableness classification "is based on the assumption that the leader has already acquired the required skills and abilities, whether by training or by experience" (clarification, p. 1) is misleading. In fact, there is nothing very special about leader training and experience. Certainly it

is true (as shown above) that inexperience and lack of training can reduce the favorableness of the situation for the leader, but so can a number of things, such as leader personality, or the inherent nature of the task to be performed. In this sense, Fiedler's new Table 1 is confusing, and his remarks concerning the "ill-fated Figure 1" of his earlier article are incorrect. In truth there was nothing ill-fated about it. It correctly showed that by increasing the favorableness of the situation you will tend to improve "fits" between leader and situation that are bad, while impairing those that are already good. The point is that this is true whether improvement of the situation occurs through experience, training, or for *any other reason*. Furthermore, it is true regardless of whether the leader is trained or untrained. In short, it does not really seem to matter why the situation is unfavorable, and, if this is the case, then inexperience and lack of training have no special significance for the Model.

For proof of this, consider Table 1 of the clarification. Take, for example, a high-LPC leader in a situation of medium favorableness. Table 1 indicates that his level of performance, assuming "adequate training and experience" will be good. The performance level for the high-LPC leader "without adequate training and experience" in a medium-favorable situation will also be good. In fact, there is not a single entry where the untrained, inexperienced leader will perform any differently than will the adequately trained leader.

The point is that the effects of lack of training or inexperience will "enter into the equation" by causing the favorableness of the situation score to be lowered. Once this happens, there is no further need to be concerned with the fact that the leader is inexperienced or untrained. That is why Fiedler's statement that the Model depends on an assumption about adequately trained and experienced leaders is a misleading one. Do lack of training and experience "enter into the equation" in some other way besides through its effect on situation favorableness? If not, his statement is wrong; if so, Table 1 is incomplete.

Fiedler extends the confusion concerning

this point on page 2 of the clarification, when he states that "a relatively inexperienced and untrained leader will perform *as if the situation were less favorable for him than for the highly trained and experienced leader*. The point is not that the untrained leader performs "as if" the situation were less favorable for him, but rather that the situation *is in fact* less favorable for him. This fact will be reflected in the untrained leader's situation favorableness score, and should not be presented as if it were a special source of concern.

#### *Is It Reasonable to Assume That Leadership Training and Experience Will Produce No Change in LPC scores?*

Fiedler (clarification, p. 4) states his assumption that leader training and experience affects performance through its effect upon situation favorableness, rejecting the "alternative hypothesis" that the increased experience and training changed the LPC scores "rather than" changing favorableness. Why is it, though, that we must choose *between* these "either-or" propositions? It seems more plausible to proceed on the assumption that such training and experience may in fact change situation favorableness, LPC scores, or both, and that we must therefore be concerned with the interactive effects of training and experience upon both the situation and the leader. In cases where "human-relations" approaches are tried upon leaders with intermediate-favorableness situations, for example, we would agree with Fiedler that leader-member relations, and therefore situation favorableness, would be likely to improve. However, we would also insist that the leader is also likely to change, probably in the direction of becoming more relationship-motivated. While Fiedler (clarification, page 4) cities studies in which subject LPC scores remained fairly stable over time, sizable changes in LPC score as a result of experience have been reported by Stinson and Tracy (1972), and Fiedler himself has cautioned that the stability of LPC scores "depends to a considerable degree on the intervening experience of the men (1967, p. 48)." It can in fact be argued that some kinds of training will not affect situation favorableness except

by first changing the motivational system of the leader.

Once we recognize that training and experience may change the leader as well as the situation, we can easily see that predicting changes in performance is a much more difficult proposition than has been suggested by Fiedler. An incomplete schema of possible changes in performance is presented in Figure 1. The horizontal arrows are identical to those in Fiedler's original Figure 1 and depict instances where training or experience alters situation favorableness without changing the leader's motivational system. Vertical arrows illustrate that training and experience may change the leader's LPC score without affecting favorableness of the situation. This would occur, for example, in cases where company policies or office politics prevent a "changed" leader from implementing his new philosophy. Finally, diagonal arrows have been included to suggest instances where both the leader and the situation are changed, and to provide a partial explanation for the failure of leadership training to systematically improve organizational performance. For example, the Contingency Model states that a task-motivated leader is ill-suited to an intermediate-favorableness situation. The horizontal arrow in Figure 1 would suggest that we institute training for such a man, so as to cause situation favorableness to improve, with resulting good performance. This recommendation is consistent with Fiedler's Table 1 (clarification, p. 7). The diagonal arrow in Figure 1, however, reminds us that we are quite likely to increase the Relationship-motivation of the leader, particularly if we resort to "human-relations" training. The result might be that the original (task-motivated leader, intermediate situation) mismatch is replaced by a new (relationship-motivated, favorable situation) mismatch, resulting in no overall change in performance.

#### *Do We Need to be Concerned with Type, As Well As Extent, of Training and Experience?*

It is likely that the *type* of training a person is exposed to is as important a consideration as whether training is instituted at all. Fiedler has combined "human relations" approaches with "the more orthodox type of

### Favorableness of the Situation

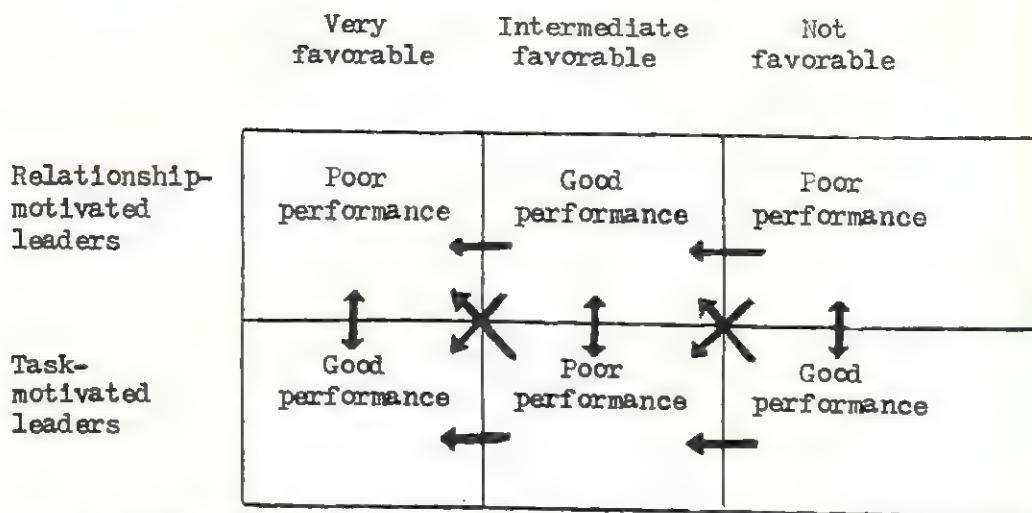


FIG. 1. Schematic representation of the hypothesized effect of training and experience on leadership performance for relationship-motivated (high LPC) and task-motivated (low LPC) leaders. (Training and experience may increase situational favorableness, leader motivation style, or both and, therefore, improve performance of some leaders but decrease performance of other leaders. Arrows indicate the predicted effect of experience and training.)

approach which is concerned with providing the leader or manager with the more technical and administrative skills . . . [1972, p. 115]." However, these types of training ought to be considered separately. The two types may be potentially able to change situational favorableness to the same extent (although this has yet to be empirically demonstrated), but they are unlikely to have identical effects on any given individual. For example, Fiedler's Hypothesis 1 (1972, p. 116) predicted that, in very favorable situations, training and experience would improve the performance of task-motivated leaders. This may be true for training which improves the task structure; it is far less likely to be true for human relations approaches. Sensitivity training, for example, will probably have a stronger effect on the LPC score of a task-motivated leader than it will upon an already very favorable situation. It may therefore be true in this case that training which improves task structure will in fact increase performance, while training of a human-relations nature will impair performance, by producing a more relationship-motivated leader for a very favorable situation.

In summary, it may well be that the Contingency Model can be of assistance in predicting the effects of leadership training and experience, and Fiedler has presented some interesting data to support his position. However, it seems to us that successful utilization of the Model for this purpose is a far more complicated proposition than it may appear to be. The problems discussed above, and probably many others as well, need to be adequately resolved if the Contingency Model is to be useful in accurately predicting the effects of leader training and experience.

### REFERENCES

- FIEDLER, F. E. *A theory of leadership effectiveness*. New York: McGraw-Hill, 1967.
- FIEDLER, F. E. Predicting the effects of leadership training and experience from the contingency model. *Journal of Applied Psychology*, 1972, **56**, 114-119.
- FIEDLER, F. E. Predicting the effects of leadership training and experience from the contingency model: A clarification. *Journal of Applied Psychology*, 1973, **57**, 110-113.
- STINSON, J. E., & TRACY, L. The stability and interpretation of the LPC score. *Proceedings of the Midwest Academy of Management*, 1972.

(Received December 29, 1972)

# MANAGERIAL SATISFACTIONS AND ORGANIZATIONAL ROLES: AN INVESTIGATION OF PORTER'S NEED DEFICIENCY SCALES

JEANNE B. HERMAN<sup>1</sup> AND CHARLES L. HULIN<sup>2</sup>

*University of Illinois*

The empirical research identifying a relationship between job satisfaction and level in the organizational hierarchy has utilized the Porter Need Satisfaction Questionnaire extensively. An attempt was made to replicate the previous findings and expand the domain of job satisfaction variables to determine the generality of the relationship. The hypothesis of different mean levels of satisfaction associated with different levels in the hierarchy was supported using the Job Description Index but was not supported using the Porter Need Deficiency Scales. The internal structure of the Porter Questionnaire and its convergence with the JDI were investigated to explore alternative explanations of the results. Characteristics of the sample and the analytic procedures of the original studies are discussed.

A series of articles by Porter in the early 1960's (1961, 1962, 1963a, 1963b, 1963c) established a new domain of attitude research relating managerial attitudes to organizational role. The original studies, which investigated differences in need satisfactions among various groups of managers, form the basis of empirical knowledge of the managerial role—job attitude relationship. Porter and Lawler's review of the research (1965) on job satisfaction across organizational levels, indicates that the domain of variables investigated in studies on organizational attitudes and organizational roles has not been much expanded since the early studies. The cumulative research results, which indicate an increasing level of job satisfaction at higher levels of the organization, are almost entirely dependent on a common set of measures (the Porter Need Satisfaction Questionnaire) and a common analytic technique (multiple sign tests). The possibility that the cumulative research may do little more than demonstrate a results-methods dependency cannot be precluded.

The research reported in this article was designed to investigate the stability of the

organizational level—job satisfaction relationship using an expanded domain of job satisfaction variables and a multivariate analytic technique.

Expansion of the domain of dependent variables is desirable in order to determine the generality of the phenomena across measures. But, since convergent validation of the Porter Need Satisfaction Questionnaire has never been published, and recent research (Imparato, 1972; Roberts, Walter, & Miles, 1971; Sonsoni & Johnson, unpublished paper) questions the validity of the five Porter scales, it seemed that expansion of the dependent variable domain without concomitant replication would not tie the extension of the research adequately to prior findings.

Results-methods dependencies also may be a function of analytic technique. A careful view of the data presented in the studies supporting the hypothesized relationship between organizational level and job satisfaction indicates that the results are not quite so overwhelming as the conclusions would suggest (see Porter, 1963c, p. 386 and then compare ElSalmi & Cummings, 1968, or Porter, 1962, 1963b).<sup>3</sup> Whenever multiple comparisons are made on correlated dependent variables the  $\alpha$  level for each hypothesis is unknown. Trend analysis on the means of the hierarchical groups, first on the 13 Porter need

<sup>1</sup> Requests for reprints should be sent to Jeanne B. Herman, Department of Psychology, University of Illinois at Urbana, Champaign, Illinois 61820.

<sup>2</sup> The authors would like to thank Andres Inn for his invaluable aid in all phases of this study and the officials of the company involved for providing an opportunity to obtain the data reported.

<sup>3</sup> At the request of the editor, the authors deleted a more detailed review and critique of this literature.

deficiency items, then on the five derived scale scores was the general analysis procedure used in many of the supporting studies. The covariation among the items is not reported, yet the scale development procedure suggests that at least some covariation is assumed. To the extent that the dependent variables covary, the significance of mean differences will be highly overstated by the assumption of independence made in multiple significance tests.

A multivariate analytic procedure that provides an overall significance test as well as individual significance tests on the dependent variables is more appropriate. Moreover, the analysis of variance model is more rigorous than multiple sign tests on means, since it tests hypotheses on between- versus within-group variance, so that group similarities and differences in the domain of interest may be explored more fully.

#### METHODS

##### *Subjects*

Subjects were four levels of supervisory personnel from a large midwestern plant involved in the manufacturing and assembly of heavy equipment. Data were collected at weekly supervisory communications sessions over a period of two months. The data reported in this study were obtained during the third and sixth sessions. The data collection was administered by the researchers who were identified to the subjects as independent university personnel engaged in studying job attitudes of managers in a large number of midwestern companies. The researchers were not involved in the substance of the communications sessions. Subjects signed their questionnaires with their names or an identifying mark so that responses from the several sessions could be matched.

Due to work scheduling problems, not all the managers completed both sets of questionnaires. However, neither of the two overlapping but slightly different samples used in the analyses differed significantly from the total sample on any of the demographic characteristics presented in Table 1.

Instruments analyzed in this study are the Porter Need Satisfaction Questionnaire (Porter, 1962) and the Job Descriptive Index (Smith, Kendall, & Hulin, 1969).

The questionnaire designed by Porter is an operationalization of Maslow's need hierarchy theory. Managers are asked to provide three responses for each of the 13 items: (a) How much of the characteristic is there connected with your present position, (b) how much of the characteristic do you think there *should be* connected, and (c) how important is this characteristic to you? The respondents

TABLE 1  
YEARS OF EDUCATION, AGE, AND TENURE DISTRIBUTION  
OF TOTAL SAMPLE (N = 174)

Item	Frequency
Years of education	
<8	15
9-11	43
High school	83
More than high school	20
"No" answer	13
Age	
20-29	13
30-39	41
40-49	66
50-59	43
>60	1
"No" answer	10
Tenure	
<1	6
1-5	6
6-10	24
11-15	10
16-20	29
21-30	76
31-40	15
>40	1
"No" answer	7

are asked to answer these three questions for each job characteristic by circling a number on a 7-point rating scale. The rating scale is anchored so that 1 represents a low or minimum amount and 7 represents a high or maximum amount.

Thirteen need fulfillment scores are based on the "now" questions; need deficiency scores are assessed by the difference between the "should be" and "now" responses. The 13 need fulfillment and need deficiency items have been classified *a priori* into the five need categories of Maslow: security, social, esteem, autonomy, and self-actualization needs.

The dependent variables generally used in the Porter studies are the 13 need deficiency item scores and 5 need deficiency scale scores. The responses to the importance questions have been analyzed only rarely (see Porter, 1963a as an exception). Occasionally need fulfillment scores are analyzed alone (Porter & Mitchell, 1967).

The Job Descriptive Index (JDI) is a checklist that asks subjects to state whether various adjectives are descriptive of five principal dimensions of their job: the work itself, the supervisor, the pay, the promotion, and the co-workers. The JDI scales were developed, factor analytically, and are moderately correlated. The JDI has demonstrated an acceptable level of convergent and discriminant validity (Vroom, 1964).

TABLE 2  
GROUP MEANS ON THE FIVE NEED DEFICIENCY SCORES

Group	Security	Social	Esteem	Autonomy	Self-Realization	<i>N</i>
Foremen	3.20	1.94	5.65	5.43	5.45	86
General foremen	2.47	.76	3.47	3.94	3.59	17
Supervisors	2.33	1.83	5.00	4.50	5.17	12
Superintendents	2.50	1.00	5.17	3.50	3.17	6

### Analysis

In order to attempt to replicate the cumulative research results associated with the Porter Scales and to extend the domain of dependent variables to another set of validated job satisfaction measures, analysis proceeded along several lines. Discriminant analysis (Tatsuoka, 1970) was used to test the hypothesis of group differences on the dependent variables. The dimensionality of the Porter items was investigated using principal axis factor analysis, with  $R^2$  communality estimates in the diagonals and varimax and oblimax rotations (Kaiser, 1970). Standard multitrait-multimethod (Campbell & Fiske, 1959) correlations were used to test the convergence of the two sets of measures of job satisfaction.

### RESULTS

The managerial level-job satisfaction hypothesis was tested as a replication using the five Porter need deficiency scales and as a validity extension using the five JDI Scales.

The sample of managers for the need satisfaction analysis consisted of 86 foremen, 17 general foremen, 12 supervisors, and 6 superintendents. The results of the discriminant function analysis on the need deficiency scales indicated that no linear combination of the five need categories significantly discriminated among the four groups. The overall  $F$  ratio ( $F_{15/312} = .78$ ) was not significant. Table 2 presents the group means on the five need deficiency scales.

The sample of managers for the JDI analysis included 111 foremen, 21 general foremen, 15 supervisors, and 11 superintendents. This discriminant analysis resulted in one highly significant ( $p \leq .01$ ) linear function and one marginally significant function ( $p \leq .06$ ). The overall test ( $F_{15/414} = 2.5$ ) indicated the solution was highly significant.

The first discriminant vector in the JDI analysis arranged the managerial groups in hierarchical order (Table 3). The scaled loadings of the items on the discriminant vectors (Table 4) indicates that group differences are primarily on work and pay satisfaction. The second dimension of group differences seems to be picking up situationally idiosyncratic variance. The superintendents who were the highest management level in the sample were least satisfied on this linear combination of variables. The scaled item loadings (Table 4) identify co-workers as the most potent variable in the discrimination on this dimension. Table 5, the group means on the five JDI variables shows the effect very clearly. Independent analyses of other data collected at the same time, but not reported here, indicated the job attitudes of the superintendents were unique and did not fit the expected pattern. There had been some rather extensive

TABLE 3

GROUP MEANS IN THE SPACE DEFINED BY THE JDI DISCRIMINANT FUNCTIONS

Group	I	II
Foremen	21.82	21.89
General foremen	24.41	20.18
Supervisors	29.14	27.50
Superintendents	30.39	16.08

TABLE 4

SCALED LOADINGS OF THE JDI VARIABLES ON THE DISCRIMINANT VECTORS

Scale	I	II
Work	88.04	54.01
Supervision	-19.99	-62.61
Pay	43.19	6.16
Promotion	-2.37	-48.86
Co-workers	-33.72	87.09

TABLE 5  
GROUP MEANS ON THE JDI SCALES

Group	Work	Supervision	Pay	Promotion	Co-workers	N
Foremen	33.26	38.61	26.72	20.22	41.17	111
General foremen	36.10	41.81	29.14	28.00	41.62	21
Supervisors	40.93	39.53	31.87	22.40	45.00	15
Superintendents	38.64	43.82	37.27	27.45	34.64	11

personnel changes in top level of management at this plant during the six months preceding the administration of the questionnaire. It is the opinion of the researchers that the pattern of group means on the second discriminant dimension may reflect this situation.

The validity extension study supported the hypothesized satisfaction differences associated with hierarchical levels; the replication study did not. Any explanation of these contradictory results required analysis of the characteristics of the Porter Questionnaire and the relationships between the JDI variables and the Porter variables.

Initially, the dimensionality of the Porter instrument was determined. It is a necessary (but not a sufficient) condition that the 13 items measure five discriminably different dimensions if the scale scores are to be used independently in further analyses. If the 13 need deficiency scores are to be used separately, the covariation among the scores

should be minimal and the factor pattern diffuse.

The results of the factor analysis of the intercorrelations of the 13 need deficiency items indicated that the first dimension accounted for 88% of the common variance. Subsequent factors accounted for 12%, 8%, 5% and 2%, respectively, of the common variance. Factor analyses were repeated on the need, have, importance, and need deficiency weighted by importance responses. The results were similar in all cases. Such a pattern could indicate a very high degree of trait variance among the 13 items, measuring 5 levels of need satisfaction; a high degree of method variance; or both. The results do not indicate which is the appropriate interpretation. The pattern of root sizes suggests that a one-dimensional solution would be the most parsimonious interpretation and one which is consistent throughout all analyses. Nevertheless, in order to allow the five need category

TABLE 6  
VARIMAX ROTATED FACTOR LOADINGS OF NEED DEFICIENCY ITEMS

Scale	I	II	III	IV	V
Security	.31	.37	.31	.21	.34
Social	.11	.12	.24	.46	.06
Social	.13	.52	-.02	.13	.04
Esteem	.67	.17	.22	.16	-.00
Esteem	.60	.10	.38	.21	.25
Esteem	.27	.49	.02	.44	-.06
Autonomy	.64	.17	.22	.16	-.00
Autonomy	.19	.11	.48	.21	.02
Autonomy	.21	.54	.49	.22	-.03
Autonomy	.36	.22	.27	.42	.08
Self Realization	.44	.49	.30	.31	.19
Self Realization	.10	.59	.37	.00	.10
Self Realization	.34	.06	.51	.16	.13
% common variance accounted for by each factor	31%	28%	22%	15%	4%

TABLE 7  
CORRELATIONS BETWEEN THE JDI SCALE SCORES AND THE NEED DEFICIENCY SCORES

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1. Work (JDI)	—	.49	—	.26	.30	—	.47	.49	—	.22	—	.03	—	.29	—	.12	—	.57	—
2. Supervision (JDI)																			
3. Pay (JDI)																			
4. Promotion (JDI)																			
5. Co-workers (JDI)																			
6. Security																			
7. Social																			
8. Social																			
9. Esteem																			
10. Esteem																			
11. Esteem																			
12. Autonomy																			
13. Autonomy																			
14. Autonomy																			
15. Autonomy																			
16. Self Realization																			
17. Self Realization																			
18. Self Realization																			

Note. N = 158 for JDI with JDI correlations, r = .13, p ≤ .05, N = 121 for Porter Items with Porter Items correlation, r = .16, p ≤ .05; and N = 105 for cross-instrument correlation, r = .18, p ≤ .05.

interpretation of the 13 items the greatest opportunity for support, the first five factors were rotated to both varimax and oblimax criteria. Since the first scale category, security, contains only one item, it is unlikely that it would identify an independent factor. But, if the factor analytic solution is indicating a method variance factor, rotating five factors should maximize the possibility that items in the four higher level need categories would converge appropriately and the method variance common to all items would form a separate dimension. Table 6 presents the varimax rotated item loadings. The items are grouped by category in the table, but the categories do not define the dimensions. Rather it appears that the loadings of items within a need category are randomly distributed across the matrix. Item loadings also do not identify one common method factor. The results of the oblimax rotation were similar.

To guard against the possibility that these results were reflecting a situationally specific phenomenon, the same analyses were done on a sample of hospital personnel (Sonsoni & Johnson, unpublished paper) and on two other samples of managerial personnel, one from a government agency, the other from a large retail corporation.<sup>4</sup> In all cases the relationships among the 13 items for each of the five responses (have, need, importance, need deficiency, and need deficiency weighted by importance) were best approximated by a one-dimensional solution.

Table 7 indicates that the degree of convergence between the Porter need deficiency items and the five JDI scales is minimal. While a number of cross instrument correlations are significant, the heterotrait correlations for both the JDI scales and Porter items are too large relative to the monotrait correlations to allow any statements about convergence. Since the JDI has been shown to converge with other measures of job satisfaction (Smith, Kendall, & Hulin, 1969) and the Porter instrument has not (Evans, 1969; Roberts, Walter, & Miles, 1971; Sonsoni & Johnson, unpublished paper), it would seem that the domain of job satisfaction identified

by these two instruments is rather heterogeneous.

## DISCUSSION

The managerial level-job satisfaction hypothesis failed to replicate on the need satisfaction scales but found support with the JDI variables. Analysis of the dimensionality of the Porter questionnaire indicated the 13 items could not support the five scale scores; but that a single dimension would be the most parsimonious use of the Porter items. With these results in mind, multivariate differences of managerial groups on the five dependent need deficiency variables would not be expected. The analytic technique, however, does not preclude significant group differences on a single linear combination of the five variables. Yet, there were no significant differences between managerial groups measured by the Porter questionnaire.

Managerial groups were significantly different in two dimensions in the JDI analyses, clearly supporting the hierarchical level-job satisfaction hypothesis. These results stand along with some other recent research (Herman & Hulin, 1972) as an independent validity extension of the cumulative research summarized by Porter and Lawler (1965). Job satisfaction of managers is related to their level in the organizational hierarchy.

Substantiation of the hypothesis without replication of previous results leaves several questions unanswered. A thorough critique of the earlier studies is beyond the scope of this article. The analytic technique used in these early studies has already been commented on. In addition, the sample on which four of the five studies (Porter 1962, 1963a, 1963b, 1963c) was based deserves special scrutiny. Seventy-six percent of the first and second level supervisors had a college degree. This percentage was not different from the percentage of the upper levels of management who were college graduates. When normally correlated variables are orthogonal due to nonrandom sample selection, generalizability is limited to other samples where years of education and level in the supervisory hierarchy are uncorrelated. Such variables are generally correlated in industrial samples.

Failure to replicate using the Porter varia-

<sup>4</sup> The authors would like to thank E. E. Lawler who very thoughtfully provided these data.

bles could be due to analytic technique or sample differences, but then the JDI analysis should not have been significant. The key is more likely the low levels of covariance between the JDI and the Porter items. Since the two questionnaires did not converge in this sample, it is quite reasonable that one set of measures would demonstrate significant group differences and the other would not. There were, however, no a priori reasons to believe that one set of measures in the job satisfaction domain would be more sensitive to group differences than the other. The lack of convergence and failure to replicate casts doubt on the conclusions about job satisfaction drawn from the research on the Porter Need Satisfaction Questionnaire. It is not the point of this discussion to discredit the validity of the hierarchical level-job satisfaction hypothesis, only to question the support for that hypothesis in the need satisfaction studies.

#### REFERENCES

- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, **56**, 81-105.
- ELSALMI, A. M., & CUMMINGS, L. L. Managers' perceptions of needs and need satisfactions as a function of interactions among organizational variables. *Personnel Psychology*, 1968, **21**, 465-478.
- EVANS, M. G. Convergent and discriminant validities between the Cornell Job Descriptive Index and a measure of goal attainment. *Journal of Applied Psychology*, 1969, **53**, 102-106.
- HERMAN, J. B., & HULIN, C. L. Studying organizational attitudes from individual and organizational frames of reference. *Journal of Organizational Behavior and Human Performance*, 1972, **8**, 84-108.
- IMPARATO, N. Relationship between Porter's Need Satisfaction Questionnaire and the Job Descrip-
- tion Index. *Journal of Applied Psychology*, 1972, **56**, 397-405.
- KAISER, H. F. A second-generation little jiffy. *Psychometrika*, 1970, **35**, 401-416.
- PORTER, L. W. A study of perceived need satisfaction in bottom and middle management jobs. *Journal of Applied Psychology*, 1961, **45**, 1-10.
- PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *Journal of Applied Psychology*, 1962, **46**, 375-384.
- PORTER, L. W. Job attitudes in management: II. Perceived importance of needs as a function of job level. *Journal of Applied Psychology*, 1963, **47**, 141-148. (a)
- PORTER, L. W. Job attitudes in Management: III. Perceived deficiencies in need fulfillment as a function of line versus staff type of job. *Journal of Applied Psychology*, 1963, **47**, 267-275. (b)
- PORTER, L. W. Job attitudes in management: IV. Perceived deficiencies in need fulfillment as a function of size of company. *Journal of Applied Psychology*, 1963, **47**, 386-397. (c)
- PORTER, L. W., & LAWLER, E. E. Properties of organizational structure in relation to job attitudes and job behavior. *Psychological Bulletin*, 1965, **64**, 23-51.
- PORTER, L. W., & MITCHELL, V. F. Comparative study of need satisfactions in military and business hierarchies. *Journal of Applied Psychology*, 1967, **51**, 139-144.
- ROBERTS, K. H., WALTER, G. A., & MILES, R. E. A factor analytic study of job satisfaction items designed to measure Maslow need categories. *Personnel Psychology*, 1971, 205-220.
- SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally, 1969.
- SONSONI, F., & JOHNSON, B. A critical evaluation of Porter's job satisfaction questionnaire. Unpublished manuscript.
- TATSUOKA, M. M. Discriminant analysis: The study of group differences. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.
- VROOM, V. *Work and motivation*. New York: Wiley, 1964.

(Received September 29, 1971)

## EQUITY THEORY AND CAREER PAY: A COMPUTER SIMULATION APPROACH

PAUL C. NYSTROM<sup>1</sup>

*School of Business Administration, University of Wisconsin-Milwaukee*

Computer simulation was used to convert Jaques' theory of equitable payment into the composites model utilized by the Haire, Ghiselli and Gordon study of career pay. Salaries of 100 subjects were stochastically allocated for 25 time periods. A Markovian process model produced a set of pay parameters that more closely replicated past empirical findings than the parameters produced by an independent process model. Distributing pay increases according to differentially developing work capacity curves yielded pay increases distributed at random with respect to past salaries. Thus, Jaques' theory of equitable payment provides one explanation for the empirical findings generated by previous studies of career pay curves.

As an employee receives pay increases over time, these pay increases accumulate into that individual's career pay curve. According to Opsahl and Dunnette (1966), "the concept of equity applies to pay-curve comparisons as well as wage comparisons, and this is an important area for investigation [p. 103]." Similarly, Weick (1966) has suggested that theories of equitable pay for work may be improved by incorporating the time dimension. However, recent literature reviews (Goodman & Friedman, 1971; Pritchard, 1969) of the substantial body of research on Adams' (1963) theory of inequity illustrate the general absence of studies concerning equity over extended time intervals and equity in permanent employment relationships. Two exceptions that do explicitly consider the time dimension are Jaques' (1961) theory of equitable payment and the research by Patchen (1961) on wage comparisons. The continuing study of merit increase problems (Giles & Barrett, 1971; Zedeck & Smith, 1968) indicates the importance of developing equitable career pay systems.

The purpose of this study was to determine whether Jaques' theory of equitable payment provides a plausible explanation for the empirical findings of a major study of career pay conducted by Haire, Ghiselli, and Gordon (1967). In order to compare Jaques' theory with the Haire et al. findings, computer

simulation techniques were developed to represent the salary allocation process within an organization.

### *Equitable Payment*

Jaques' theory of equitable payment relates three variables; level of work (W) in the role occupied, level of work capacity (C), and level of payment (P). Most of the research and criticism (Hellriegel & French, 1969) has focused on time span of discretion as a measure of the level of work variable. It is ironic that so little attention has been given the Capacity variable, considering its major role in the social comparison process:

Having related ourselves to our work, we are in a position to compare our own job with other jobs, not as we may think by means of job comparisons, but by means of comparisons of our own capacity with that of our friends and associates. I find myself forced to the conclusion that there is great precision in our ability to compare levels of capacity in one another. I think it is done by myriad clues of the way in which the other person talks and thinks, in particular the way in which he organizes his perceptions [Jaques, 1961, p. 224].

It is hypothesized (Jaques, 1956) that capacities for work develop in regular patterns over time; that working in a role equivalent to one's work capacity is experienced as a state of psychological equilibrium; and, finally, that employees seek jobs with levels of work consistent with their current capacity for work. Again, consider Jaques' formulation of the

<sup>1</sup> Requests for reprints should be sent to Paul C. Nystrom, School of Business Administration, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201.

equity process:

. . . a person's primary drive with respect to his work is towards a level of work that can absorb his capacity; towards a job in which he can use his capabilities to the full. His drive for money follows from this prime urge to employ his skill and talent in his work—and is a drive for a rate of reward that is equitable and gives him a relative economic status coinciding with his capacity [Jaques, 1961, p. 186].

Jaques observed that salary changes seemed to be processed by employees as movements towards or away from an internal standard. This internal standard is the rate of progress expected by the employee. After plotting the earning histories for 250 employees of five companies relative to their ages, Jaques fitted a smoothed set of curves designated as Standard Earning Progressions (SEP). Plotting on semilog paper clearly demonstrates that the most rapid rates of progress are made in the earlier working years; the curves are described as following the sigmoidal progression characteristic of biological growth (see Figure 1). The SEP curves are a hypothetical construct representing development of individuals' work capacities for exercising discretion. A person's capacity curve is inferred from a study of past instances where that person had performed effectively and had experienced satisfaction with earnings. Following Jaques, the SEP curves were regarded by the author as an array of overlapping bands, because this interpretation recognized the imprecisions involved in the initial determination of these SEP curves.

A limitation of the present study is that all three variables related by Jaques' theory are not observed. Recall that equity was purportedly the outcome of a dynamic process with a sequence from Capacity → Work → Pay. If one assumes that subjects receive pay that is equitable for the level of work performed, then the problem becomes underemployment [ $C \geq (P = W)$ ] or overemployment [ $C \leq (P = W)$ ]. If one does not assume  $P = W$ , then there is no way to distinguish between the twelve different patterns of inequity enumerated by Jaques (1961).

#### Career Pay Curves

As mentioned earlier, an individual's present pay level is composed of the past pay level

plus any pay increases received during the period. In their pioneering study of empirical career pay curves, Haire et al. (1967) recognized this cumulative property by using the statistical model for composites. The composites model utilized by Haire et al. is summarized in three equations:

- Let:
- $k_s$  = pay for subject  $s$  at year  $t$
  - $i_s$  = pay increase for subject  $s$  during year  $t$
  - $r_{k,k+i}$  = correlation of pay at year  $t+1$  with pay at year  $t$
  - $r_{ki}$  = correlation of pay at year  $t$  with pay increases allocated during year  $t$
  - $\sigma_i$  = standard deviation of pay increases allocated during year  $t$
  - $\sigma_k$  = standard deviation of pay at year  $t$
  - $\sigma_{k+i}$  = standard deviation of pay at year  $t+1$

$$\text{Then: } (k+i)_s = k_s + i_s \quad (1)$$

$$\sigma_{k+i} = \sqrt{\sigma_k^2 + \sigma_i^2 + 2\sigma_k\sigma_i r_{ki}} \quad (2)$$

$$r_{k,k+i} = \frac{\sigma_k + \sigma_i r_{ki}}{\sigma_{k+i}} \quad (3)$$

A major research question concerns the decision rule for allocating pay increases to a group of employees. For example, the largest raises can be allocated to those employees with the highest salaries. At the other extreme, the largest raises can be allocated to those employees with the lowest salaries. In fact, pay increases are often distributed at random with respect to past salaries (Haire et al., 1967), leading these authors to state that:

The approach of  $r_{ki}$  to 0 has disturbing psychological implications. Baldly, it means that raises are randomly distributed with respect to performance . . . Psychologists are seldom in a position to say how data *ought* to be in the real world. Here it seems possible. The correlation  $ki$  *ought* to be positive and significantly different from zero [p. 15].

And, again, "If raises are randomly distributed with respect to past salaries, consistent striving seems pointless [Haire, 1965, p. 16]." Similarly, "If there is no contingency between behavior and money, the individual manager cannot be expected to respond as if there were

[Campbell, Dunnette, Lawler, & Weick, 1970, p. 368]." An alternative explanation for the observed low correlations would be that job performance is being rewarded, but one must then relax assumptions that skills are acquired predictably and that job performance is consistent over time (Opsahl & Dunnette, 1966). Ghiselli (1965) advanced four reasons why  $r_{ki}$  might equal zero even though the intent is to reward performance; (a) performance may vary randomly, (b) performance criteria may vary between years, (c) employee mobility may be between jobs with different requirements, or (d) performance may not be reliably measured.

In a related study, Brenner and Lockwood (1965) reported finding that salary at one date was a good predictor of salary at a later date. Further, they reported finding that the predictability of salary levels improved with tenure. Brenner and Lockwood were disturbed by these high correlations ( $r_{k,k+i}$ ), whereas Haire et al. expressed concern over declining correlations ( $r_{k,k+i}$ ) over time. Both studies' implications are speculative, neither study having examined the psychological outcomes of salary distribution parameters. Indeed, existing research generally has not related attitudinal data on perceptions regarding pay to hard data on compensation dollars (Hinrichs, 1969). Career pay curves reflect several aspects of an organization's compensation practices, and there is some evidence suggesting that compensation practices influence employee work attitudes. For example, Mahoney (1964) found that managers' compensation preferences closely paralleled their perceptions of current compensation practices. Hinrichs also investigated salary practices as a factor shaping perceptions of pay, concluding that the current level of earnings was an important variable affecting employee perceptions regarding pay increases.

#### *Stochastic Process Models*

It is not reasonable to assume that organizations have yet achieved a utopian situation in which each employee is always in a state of equilibrium between capacity, work, and payment. Even assuming that work capacity develops smoothly, there are several organizational factors causing both payment and work

performance to deviate from this smoothly developing work capacity curve. The level of work required in a role is likely to change in discrete steps over time, rather than as a smooth function. Job mobility introduces additional discontinuities. Compensation practices tend to emphasize periodic reviews and pay increases, thereby functioning as another source for discrete steps in career pay curves. The accumulation of salary increases over time also tends to evoke the organizational response of job reevaluation and expansion of the salary structure (Nystrom, 1970). Thus, several organizational factors are expected to contribute to a pattern of deviations between actual pay and equitable payment for work capacity. Note that one can continue to assume that the organization is intending to reward job performance.

These institutional discontinuities were simulated by utilizing two finite stochastic process models, an independent process and a Markov process, to generate patterns of movement between SEP curves. An employee receives a salary associated with one of the nine SEP curves included in the sample (see Figure 1). The curve to which the employee is allocated is the state occupied at a specific time or stage in the process. Patterns of movement between states over time are represented by transition probabilities; three different probability distributions were studied. The purpose of the stochastic process models was to simulate the rather small discrepancies between work capacity and salary described in Jaques' (1961) case studies.

#### METHOD

Briefly, the research methodology involved generating statistical parameters describing salary distributions for a sample of subjects paid in a manner paralleling Jaques' developmental curves. The subjects were drawn such that they replicated one sample in the Haire et al. study. Then, these subjects were moved on and between the SEP curves over time by stochastic process models embedded in a computer simulation program. Finally, the sensitivity of results to alternative input parameters was examined. Each of these stages in the research methodology will now be discussed in greater detail.

*Data conversion.* One methodological problem concerned the conversion of Standard Earnings Progression curves to inputs compatible with the statistical model of composites. Both Jaques and Haire et al. plotted data on two-dimensional arrays, with a time measure on the horizontal axis and an income measure on the vertical

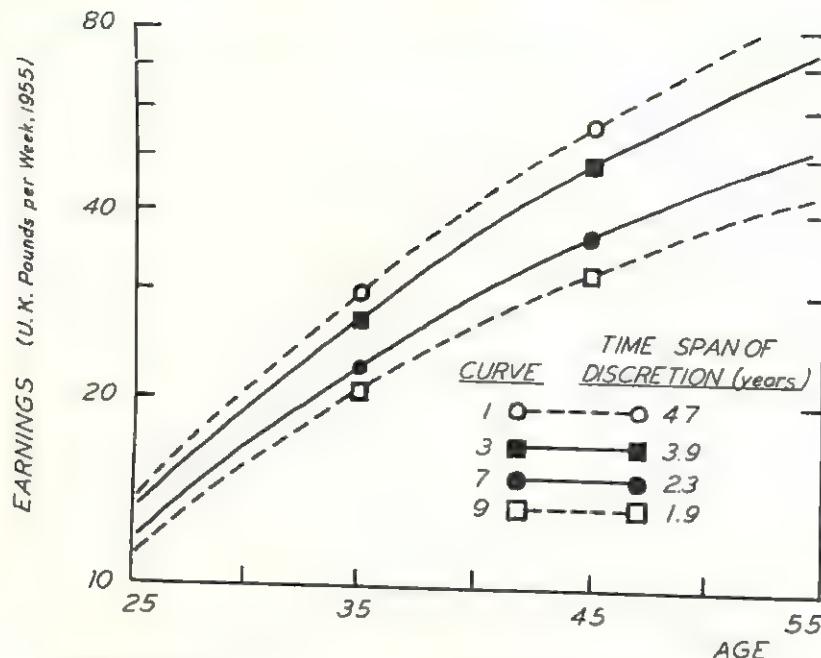


FIG. 1. Four of nine standard earning progression curves in sample.

axis (Figure 1). The Haire et al. sample identified as Company B was selected because of its middle region within the SEP array of curves. Company B was composed of up to 89 executives in jobs at the \$20,000 to \$22,000 range of annual incomes.

The horizontal axis of the Haire et al. approach was the 25-year period from 1938 to 1962, while the Jaques approach used a horizontal axis of ages from 20 to 65. An assumption must be made to convert from ages to calendar years. Managers in Company B had at least twenty years of service. It was assumed that managers joined the firm at age 25. The relevant age range was then between 45 years of age (20 years of service) and 50 years of age (given that 25 years is the duration of this study).

Turning to conversion of the vertical axis, U. K. pounds sterling were converted to dollars following Jaques (1961, p. 294). The SEP curves were reported for incomes in 1955, necessitating a correction for changes in general levels of wages when one is studying other calendar years. Deflating the salary range of \$20,000-\$22,000 in 1962 by a correction factor (.77) yielded a comparable income of approximately 42.3 to 46.5 pounds per week. Thus, the applicable region was from the SEP curve for time span of discretion level of 3.9 years (\$22,000 at age 45) to the SEP curve associated with a time span of 2.3 years (\$20,000 at age 50). Boundary SEP curves had time spans of 4.7 years and 1.9 years.

**Simulation.** A second methodological problem concerned the simulation of salary allocations as stochastic processes. A computer program was developed to simulate the movement of 100 individual employees over a period of 25 years.

In Model 1, an independent process, each of the five

central SEP curves in the sample is a track onto which a group of employees ( $n_m$ ) is initially assigned. For each of the five central SEP curves, there is an associated vector of transition probabilities. Each vector has five elements;  $\pi_m = (\rho_{m+2}, \rho_{m+1}, \rho_m, \rho_{m-1}, \rho_{m-2})$ . Thus, it was necessary to add two boundary curves onto each end of the sample of curves. Multiplication of each element of the employee distribution vector ( $n_m$ ) by its associated transition probability vector ( $\pi_m$ ) yields the employee distribution among SEP curves at time  $t + 1$ .

In Model 2, a Markovian process, the salary at time  $t + 1$  becomes a function of the salary paid at time  $t$ . It is assumed that this stochastic process is of order one, so that the probability of occupying state  $j$  at stage  $t + 1$  is conditional only upon the state  $i$  occupied at stage  $t$ . It is also assumed that the parameters of the process do not vary over time, so that the process is stationary. The vector distribution of employees ( $n_m$ ) is multiplied by the salary allocation process as represented by transition probabilities in a square matrix  $P$  (9 × 9). In Model 2, an employee can ultimately be compensated according to any of the nine salary curves under study. An employee does not track along a specific SEP curve, as in Model 1, but movement remains limited to a specified range of curves at any one stage in the stochastic process.

Movement of individuals between salary curves from one time period to the next was governed by a random number generator. A generated random number was compared with the cumulative frequency associated with the particular probability distribution under study, thereby stochastically determining the new salary level for the individual employee. Subtracting the previous salary from the new salary yielded the amount of salary increase. After completing this procedure for all 100

employees for one time period, the computer program then calculated variables  $\sigma_i$ ,  $\sigma_{k+i}$ ,  $r_{ki}$ , and  $r_{k,k+i}$  for that period.

*Sensitivity analysis.* In order to investigate the generalizability of the findings, both the independent process (Model 1) and the Markovian process (Model 2) were run with three transition probability distributions. Probability distributions studied were a rectangular distribution of equal probabilities (.2, .2, .2, .2, .2), a normal curve distribution (.0617, .2445, .3776, .2445, .0617), and a bimodal distribution (.05, .35, .20, .35, .05). The bimodal distribution approximated the butterfly-shaped distribution of optimal incongruity in the hypothesized relationship between adaptation level and positive affect (Hunt, 1965).

In Model 2, the above probability distributions described the process for the central SEP curves under study (curves 3-7). The boundary SEP curves (1, 2, 8, 9) were modeled in a manner similar to a random walk which is partially reflected at both boundaries (Kemeny & Snell, 1960).

## RESULTS

Modeling the allocation of salaries as a stochastic process produced pay parameters from Jaques' data that are very similar to those reported in the Haire et al. study. Therefore, Jaques' theory of equitable payment provides one potential explanation for previously observed career pay parameters. The computer simulation outputs were compared with the earlier studies in terms of the behavior of the processes under study. Major variables for comparison were the consistency of the allocations of pay increases relative to previous salary levels ( $r_{ki}$ ), the changing scope of aspirations over time ( $\sigma_{k+i}$ ), and the probability of status changes ( $r_{k,k+i}$ ).

It is obvious, from the composites formulae mentioned earlier, that the correlation of pay with raises ( $r_{ki}$ ) is a key pay parameter. Correlations ( $r_{ki}$ ) produced by the six computer simulations were predominantly negative; only 14 of the 150 correlation coefficients were positive in sign, and the 150  $r_{ki}$  ranged from +.17 to -.53. Average correlations for each model version over a twenty-five period simulated history are reported in Table 1. The Fischer's transformation method employed by Haire et al. was used to calculate all of the average correlations reported. When comparing the salary increase decision rule ( $r_{ki}$ ), the Markovian model simulations produce small negative coefficients (-.10, -.14, -.24) corresponding closely to the Haire et al. finding

TABLE 1  
COMPARISON OF COMPUTER SIMULATION OUTPUTS  
WITH EMPIRICAL FINDINGS<sup>a</sup>

	Average correlation of pay with pay increase ( $r_{ki}$ )	Average correlation of pay with previous pay ( $r_{k,k+i}$ )	Standard deviation of salaries in year 25 ( $\sigma_{k+i}$ )
Simulations of 25 periods duration			
Model 1: Independent			
Rectangular	-.42	.51	\$2,630
Normal	-.30	.68	2,439
Bimodal	-.32	.65	2,366
Model 2: Markovian			
Rectangular	-.24	.74	3,062
Normal	-.14	.83	2,780
Bimodal	-.10	.84	3,038
Company B (1954-58) <sup>a</sup>	-.18	—	2,500

<sup>a</sup> The findings reported here are from the Haire, Ghiselli, and Gordon (1967) study.

(-.18). The sensitivity of  $r_{ki}$  to different allocation decision parameters is demonstrated by comparisons between simulation runs within each model. Thus, even within a homogeneous sample of employees, small deviations between actual salaries and SEP curves representing work capacities result in low and negative correlations between past salaries and pay increases.

Simulation runs produce a smooth pattern of increasingly large standard deviations ( $\sigma_i$  and  $\sigma_{k+i}$ ) over time. A similar smooth pattern was observed by Haire et al. Standard deviations in the terminal period are reported in Table 1.

Another process behavior of interest concerns the predictability of pay by pay over an increasing number of time periods. Results indicate that the longer the time interval, the greater the probability of status changes among employees. Declining predictability of pay by pay over lagged years is apparent in Table 2. Only the Markovian model outputs are relevant here; an independent process produces similar correlations ( $r_{k,k+i}$ ) at each stage, and these were reported in Table 1. The criterion or base period for the Brenner and Lockwood data is 20½ years of seniority, for the Haire et al. Company B data is 1958, and for the Model 2 simulation data is period 21 or

TABLE 2  
CORRELATIONS OF PAY WITH PAY ( $r_{k,k+i}$ )  
OVER LAGGED YEARS

Lagged years	Model 2—Markovian process			Company B <sup>a</sup>	Brenner & Lockwood
	Rectangular	Normal	Bimodal		
1	.78	.84	.86	.94	.99
2	.51	.76	.75	.86	.95
3	.47	.68	.70	.80	.94
4	.41	.67	.67	.74	.92
5	.39	.64	.59	.69	.91
6	.30	.53	.55	.62	.88
7	.28	.45	.46	.59	.86
8	.25	.41	.44	.55	.79
9	.26	.35	.34	.41	.74
10	.13	.19	.21	.38	.68

<sup>a</sup> Company B findings are from Haire, Ghiselli, and Gordon (1967) study.

age 45. The Markovian model yielded correlations of pay with pay ( $r_{k,k+i}$ ), over lagged years, closely paralleling the empirical findings of Haire et al. All three studies reveal a smoothly declining curve of correlation coefficients ( $r_{k,k+i}$ ) over lagged years, although the Brenner and Lockwood sample did not exhibit as substantial a decline.

TABLE 3  
COMPARISON OF SALARY COSTS INCURRED  
BY ALTERNATIVE MODELS

Stochastic process model	Total salary costs ( $k$ ) for 100 employees over 25 periods		
	First period	Last period	All periods
1: Independent			
Rectangular	\$453,540	\$1,788,400	\$26,473,320
Normal	452,960	1,780,800	26,398,410
2: Markovian			
Rectangular	455,660	1,865,100	27,043,250
Normal	455,660	1,797,700	26,292,580
Bimodal	455,070	1,830,500	26,529,840
Maximum Difference	3,170	84,300	750,670
Highest Lowest $\times 100 =$	100.7%	104.7%	102.9%

The statistical model for composites does not control the total salary cost ( $k$ ) nor the total amount of salaries allocated as pay increases ( $i$ ). Yet, cost parameters are of considerable interest to an organization. A comparison of the salary costs incurred by the alternative computer simulation models (Table 3) indicates that differences in cost were small relative to the total dollars involved.

## DISCUSSION

The general finding of this study was that allocation of salaries in a manner consistent with Jaques' data and theory of equitable payment produces career pay curves similar to those reported in the Haire et al. study. Whereas the Haire et al. study was a major empirical work describing how pay is distributed over time, Jaques' theory provides an explanation of the psychological consequences.

In particular, the allocation of pay increases at random with respect to past pay is not necessarily a compensation policy to be avoided. An  $r_{ki} \leq 0$  is not necessarily evidence that performance is unrewarded, nor that motivation is thereby reduced. Thus, the use of group parameters to make inferences about individual behavior may yield inappropriate conclusions. In summary, allocating salaries by SEP arrays which represent individual differences in the development of work capacity produces  $r_{ki} \leq 0$ . According to the theory of equitable payment, payment consistent with work capacity contributes to a psychologically desirable state of equilibrium.

There is an interesting parallel between the curves discussed in this paper and the extensive work reported by Bloom (1964) on stability and change in human characteristics. Several human attributes having a cumulative property, such as height or intelligence, exhibit a pattern of high stability over time while often exhibiting low or even negative correlations between initial measures and gains in the next time interval. Both the empirically observed career pay parameters and the hypothetical construct of work capacity development curves are consistent with much of Bloom's work.

The findings reported here do not constitute a validation of the Standard Earning Progression curves. Nor can one justifiably conclude that it was payment according to work capacity

which yielded the findings in the Haire et al. study. One can merely conclude that the portion of Jaques' theory of equitable payment concerning longitudinal development of work capacity provides one plausible explanation for some heretofore perplexing career pay parameters. In addition, this study illustrates the potential usefulness of computer simulation as a research methodology for comparing findings from different studies. A computer simulation approach to replication and construct validation avoids costs associated with data generation, and may also avoid problems associated with comparing findings from different research designs.

## REFERENCES

- ADAMS, J. S. Toward an understanding of inequity. *Journal of Abnormal and Social Psychology*, 1963, 67, 422-436.
- BLOOM, B. S. *Stability and change in human characteristics*. New York: Wiley, 1964.
- BRENNER, M. H., & LOCKWOOD, H. C. Salary as a predictor of salary: A 20-year study. *Journal of Applied Psychology*, 1965, 49, 295-298.
- CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill, 1970.
- GHISELLI, E. E. The effects on career pay of policies with respect to increases in pay. In R. Andrews (Ed.), *Managerial compensation*. Ann Arbor: Braun & Brumfield, 1965.
- GILES, B. A., & BARRETT, G. V. Utility of merit increases. *Journal of Applied Psychology*, 1971, 55, 103-109.
- GOODMAN, P. S., & FRIEDMAN, A. An examination of Adams' theory of inequity. *Administrative Science Quarterly*, 1971, 16, 271-288.
- HAIRE, M. The incentive character of pay. In R. Andrews (Ed.), *Managerial compensation*. Ann Arbor: Braun & Brumfield, 1965.
- HAIRE, M., GHISELLI, E. E., & GORDON, M. E. A psychological study of pay. *Journal of Applied Psychology*, 1967, 51 (4, Whole No. 636).
- HELLRIGEL, D., & FRENCH, W. A critique of Jaques' equitable payment system. *Industrial Relations*, 1969, 8, 269-279.
- HINRICHES, J. R. Correlates of employee evaluations of pay increases. *Journal of Applied Psychology*, 1969, 53, 481-489.
- HUNT, J. McV. Intrinsic motivation and its role in psychological development. In D. Levine (Ed.), *Nebraska symposium on motivation*. Lincoln: University of Nebraska Press, 1965.
- JAQUES, E. *Measurement of responsibility*. Cambridge: Harvard University Press, 1956.
- JAQUES, E. *Equitable payment: A general theory of work, differential payment, and individual progress*. New York: Wiley, 1961.
- KEMENV, J. G., & SNELL, J. L. *Finite Markov chains*. Princeton: Van Nostrand, 1960.
- MAHONEY, T. A. Compensation preferences of managers. *Industrial Relations*, 1964, 3, 135-144.
- NYSTROM, P. C. Corporate manpower planning: The role of policy in a model. *Proceedings of the 23rd Annual Winter Meeting of the Industrial Relations Research Association*, 1970, 278-285.
- OPSAHL, R. L., & DUNNETTE, M. D. The role of financial compensation in industrial motivation. *Psychological Bulletin*, 1966, 66, 94-118.
- PATCHEN, M. *The choice of wage comparisons*. Englewood Cliffs, N. J.: Prentice-Hall, 1961.
- PRITCHARD, R. D. Equity theory: A review and critique. *Organizational Behavior and Human Performance*, 1969, 4, 176-211.
- WEICK, K. E. The concept of equity in the perception of pay. *Administrative Science Quarterly*, 1966, 11, 414-439.
- ZEDECK, S., & SMITH, P. C. A psychophysical determination of equitable payment: A methodological study. *Journal of Applied Psychology*, 1968, 52, 343-347.

(Received October 18, 1971)

## INTERNAL-EXTERNAL CONTROL AS A PREDICTOR OF TASK EFFORT AND SATISFACTION SUBSEQUENT TO FAILURE<sup>1</sup>

HOWARD WEISS<sup>2</sup> AND JOHN SHERMAN

*New York University*

The Rotter Scale of Internal-External Control is used to predict effort expended on a task subsequent to failure on a similar task. Given that a person has an initial need for success and that he expects that working hard will result in success, it is hypothesized that, after failing the task, "Internals" will maintain their initial expectancy and expend more effort on subsequent tasks than "Externals," who will decrease their pre-failure expectancies for success. The results of this study confirm this hypothesis, but fail to support secondary predictions regarding task satisfaction.

A number of theorists have proposed that goal directed behavior is a multiplicative function of the value of the goal and the expectancy that the particular behavior will be instrumental in attaining the goal. Atkinson (1964) has used this model to explain achievement behavior and Vroom (1964) and Lawler and Porter (1967) have applied it to industrial motivation.

Vroom (1964), in particular, expands upon this "expectancy-value" relationship and delineates two potential outcomes or goal regions. Outcome 1 is a goal region deriving its valence through its instrumentality for reaching Outcome 2. Outcome 2 directly fulfills a specific need. The strength of behavior is a multiplicative function of the valence of Outcome 1 and the expectancy or likelihood that the specific behavior will result in Outcome 1.

Given a need for success, completing any particular task (Outcome 1) will have a positive valence if it is seen as being a way of satisfying that need (Outcome 2). Following the model, an individual will then work hard on the task to the extent that he has an expectancy that hard work will lead to task completion. Two workers, equal in their need

for success will differ in task effort directly with their difference in expectancy that effort will lead to completion. Evidence for the validity of this concept can be found in both industrial (Galbraith & Cummings, 1966; Hackman & Porter, 1968; Hill, Bass, & Rosen, 1970) and nonindustrial (Arvey & Dunnette, 1970) settings.

The model, as formulated, makes no prediction as to what effect failure to reach Outcome 1 has on the individual's behavior. Feather (1966a) working within the Atkinson framework has shown that failure affects subsequent effort by lowering the expectancy that effort will lead to success. However, it is possible to expand upon Feather's conclusions and hypothesize two potential effects of the failure experience. The individual may change his original expectancy or he may reevaluate the adequacy of his original behavior by maintaining his belief that hard work leads to success. Thus if the individual has the original expectancy that hard work will lead to task completion, a failure experience may cause the individual to lower this expectancy or it may cause him to work harder.

Feather (1966b) and Rotter, Seaman and Liverant (1962) hypothesize that the concept of Internal-External Control of Reinforcement can be used to predict expectancy change subsequent to the failure experience. Rotter (1966) defines Internal-External Control as a concept that specifies an individual's perception of causality of reinforcement. Individuals who believe in Internal Control of Reinforcement (Internals) perceive reinforce-

<sup>1</sup> The authors wish to thank Joseph Weitz for his assistance in the technical and theoretical preparation of this study and for his critical review of an earlier manuscript, and Peter Hom for his assistance in the collection of data.

<sup>2</sup> Requests for reprints should be sent to Howard Weiss, Department of Psychology, New York University, 21 Washington Place, Room 300, New York, New York 10003.

ment as being contingent upon their own behavior. On the other hand, those who believe in External Control (Externals) perceive reinforcement as being controlled by factors other than their own behavior.

The present study is an attempt to incorporate the concept of Internal-External Control with Expectancy-Value theory and predict behavior subsequent to failure. A number of hypotheses were tested. Rotter (1962) states that the belief in Internal or External Control holds true for both positive and negative reinforcement. Thus, after a failure experience (negative reinforcement), Internals would attribute causality to their own behavior while Externals would attribute causality to factors other than their own behavior. For Internals, failure should cause them to reinterpret the adequacy of their original behavior while the same failure experience should cause Externals to lower their original expectancy. Therefore, we hypothesize that given a chance to repeat a task on which they have failed Internals will work harder than Externals (Hypothesis 1).

A second hypothesis, tested in the present study, derives from statements by Rotter (1962) and Hersch and Scheibe (1967) that a person who believes in External Control may expect either success, failure, or no consistency in reinforcement. Similarly, Internals can have either success or failure expectancies by having either high or low self esteem. Therefore, there is no reason to believe that Internality or Externality is related to the original expectancy of task success, and, therefore, no reason to believe there will be any difference in original effort (Hypothesis 2). To the extent that Hypothesis 2 is true, any relationship found between Internal-External Control and task effort following the failure experience can be attributed to the effect of the failure experience.

In addition, differences in belief in Internal or External Control of Reinforcement may be related to task dissatisfaction after failure. A number of researchers (Herzberg, Mausner, & Snyderman, 1959; Kuhlen, 1963; Vroom, 1964) have stated, although in different ways, that task satisfaction is related to the degree of need fulfillment the task provides. It is therefore our third hypothesis

that Externals, who are more likely to attribute failure causality to the task, should be more task dissatisfied than Internals, who are more likely to attribute causality to their own efforts (Hypothesis 3a).

Although the literature shows no necessary relationship between effort and job satisfaction, situational factors may influence the presence or absence of such a relationship (Katzell, Barrett, & Parker, 1961). In this situation, by combining Hypothesis 3a with Hypothesis 1 an alternate hypothesis that task dissatisfaction should be related to effort can be derived (Hypothesis 3b).

## METHOD

### *Subjects*

Forty-one male undergraduate students enrolled in the introductory psychology courses at New York University served as subjects, in partial fulfillment of their course requirements.

### *Procedure*<sup>3</sup>

To make salient the need for success each subject was initially informed that he would be participating in a study of intelligence and related psychological variables. In addition, he was told that since the particular test under study was designed for the general population, he, as a college student, should perform well on it. To insure that successful performance would not be taken for granted the subject was told that his performance would depend upon the effort he expended.

A preliminary test consisting of 24 interconnected mazes (scored by the number completed within 4 minutes) was then administered. In full view of the subject his pretest was corrected for errors and the results were entered into a "normative" distribution of other subjects' pretest scores. The distribution was constructed such that the subject's performance appeared to be well above average. This procedure was introduced to give the subject confidence with the test materials (it was anticipated that some people would doubt their ability to deal with maze-type materials) and to reinforce the expectancy that given hard work he would succeed on the intelligence test.

Prior to the "intelligence test" the subject was asked to answer the question "If you try to do your best, how well do you think you will do on the next maze?" by placing a check anywhere along a 7-point scale with endpoints marked "1" (poor) and "7" (excellent). The differences between this and a subsequent response to the same question

<sup>3</sup> Interested readers should write to the authors for a more complete description of methodology (see address in Footnote 2).

TABLE 1  
VARIABLE INTERCORRELATIONS

	Length of stay on Maze B	Pretest	JDI	Expectancy 1	Expectancy 2	Expectancy 2 - Expectancy 1
I-E Control	-.43**	.04	.00	-.15	-.03	.07
Length of stay on Maze B		-.38*	.09	.01	-.03	-.03
Pretest			-.48**	.23	-.08	-.22
JDI				-.24	.03	.17
Expectancy 1					-0.2	-.47**
Expectancy 2						.78**

\*  $p < .05$ , two-tailed.

\*\*  $p < .01$ , two-tailed.

later in the experiment was used as an indication of changes in expectancies for success.

The subject was given four minutes to work on the "intelligence test," an unsolvable maze (A) which was large enough to confuse the subject and prevent him from discovering it was unsolvable.<sup>4</sup> The purpose of this procedure was to give each subject the experience of failure on the experimental task. When the allotted time for Maze A expired the subject was given a modified form of the "Work" section of the JDI (Cornell Job Description Index, Smith, et al., 1969). This instrument was intended as a measure of the subject's satisfaction with the task. Upon completing the JDI the subject was informed that, contrary to the results of the pretest, he had not done well on the intelligence test; however, since the test was still in the "experimental" stage, he would have a chance to work for an unlimited amount of time on an alternate form of the test (Maze B) he had just failed. To determine whether his expectancies for success had changed or had been maintained the subject responded to the same expectancy question he answered just prior to working on Maze A.

Maze B, also unsolvable, but much larger and more intricate than Maze A, was administered without a time limit.<sup>5</sup> The length of time a subject persevered at this task was used as an index of effort.

The subject was debriefed on the experiment to this point and the Rotter Scale of Internal-External Control of Reinforcement (I-E scale) was administered. High scores on this instrument indicate belief in External Control, while low scores indicate belief in Internal Control of Reinforcement (Rotter, 1966). The I-E scale was introduced as a questionnaire being studied by another experimenter who was sharing the experimental time. The subject was told that scores on this final questionnaire might

<sup>4</sup> Any subject suspecting at this or any other point during the experimental session that any maze was unsolvable was excluded from the data analysis.

<sup>5</sup> A 52-minute cutoff was established so the experiment could be completed within the allotted subject time.

later be related to the results of the experiment he had just participated in.

## RESULTS

The intercorrelations among the six experimental variables for the total sample are reported in Table 1.

As predicted by Hypothesis 1, those who score low on the Rotter Scale, and are therefore more Internal, stay longer on Maze B. That this result is due to the failure experience and not to any original effort difference between Internals and Externals can be seen by the confirmation of Hypothesis 2; no relationship exists between Internal-External Control and scores on the pre-manipulation pretest. In addition, the expectancies that effort will lead to success are not significantly different between Internals and Externals. (The correlation between I-E scale and Expectancy Check No. 1 was  $-0.15$ ,  $p > 0.05$ .)

The results of the Expectancy Check do not, however, lend support to Hypothesis 1 since the correlation between the I-E scale and Expectancy Check No. 2 minus Expectancy Check No. 1 is  $0.07$  ( $p > 0.05$ ).

The zero correlation between the JDI and the I-E scale, and the low correlation between the JDI and length of stay give no support to Hypotheses 3a and 3b.

It is also evident, although neither predicted nor surprising, that ability as measured by the pretest is inversely related to length of stay or effort. Furthermore, an unpredicted yet significant relationship exists between the pretest and JDI; higher ability and motivational levels are associated with lower task satisfaction.

## DISCUSSION

In this experiment, the concept of Internal-External Control of Reinforcement has been shown to be related to the behavior which follows a failure experience. The hypothesis that Internals will see failure as being caused by their own behavior (and therefore keep original expectancy that effort leads to success) while Externals will see failure as being caused by factors other than their own behavior (and therefore change their original expectancy) has been supported by the behavioral criterion. After the failure experience, Internals work harder than Externals. This hypothesis was not, however, supported by self-reports of expectancy before and after the failure experience. It was believed that the change in expectancy would be reflected by differing Expectancy Check differentials between Internals and Externals. This did not occur. However, in light of the strong behavioral support for the hypothesis, it is logical to conclude that the check rather than the manipulation was at fault. In addition, it is the authors' belief that a behavioral rather than a self-report criterion should be the goal of an experiment such as this one. The present study was designed with this consideration in mind. (For a more adequate discussion of the problems involved in self-report techniques, see Kiesler, Collins, & Miller, 1969.)

The lack of relationship between the I-E scale and the pretest supports the second hypothesis and thereby lends credence to the interpretation of the results according to Hypothesis 1. However, since the pretest is not a pure measure of the general tendency to expend effort or to have maze ability, an alternative hypothesis consistent with the I-E scale-pretest correlation may be offered to explain the relationship between the I-E scale and length of stay. Given that the pretest measures both effort and ability, if Internals have lower ability and tend to expend more effort than Externals, the obtained zero correlation between I-E scale and pretest could be explained. However, studies cited by Rotter (1966) and Ewen (1971)<sup>6</sup> indicate that there is no relationship between the I-E scale and ability.

Job satisfaction was shown not to be related to either the I-E scale or length of stay (effort). It was originally hypothesized that Externals would view the task as being the cause of failure and therefore be more job dissatisfied than Internals. This was not supported by the data. However, both hypotheses were originally stated in terms of job dissatisfaction. The data show that most subjects were satisfied with the tasks, even after failure (mean JDI = 26.93).

There are at least four tenable explanations for the failure to confirm the hypotheses. The first, of course, is that need gratification theories of job satisfaction are incorrect. However, three other factors preclude any wholesale rejection of these theories. The task of solving mazes is usually found to be interesting and enjoyable by students; the demand characteristics of the experimenter may have caused the subjects to feel compelled to rate the task as interesting, and although the concept of Internal-External Control predicts that Externals will look to external factors to explain their failure, it does not specify what these factors will be. Factors other than the task could have been seen by Externals as being responsible for their failure.

The significant negative relationships between the pretest and JDI and between the pretest and length of stay, while unpredicted, are important in terms of the hypothesized relationships among the I-E scale, length of stay and JDI. These findings suggest that the effects of the pretest should be held constant in each experimental variable. The correlation between the I-E scale and length of stay (holding pretest constant) was  $-0.45$  ( $p < 0.01$ , two-tailed); thus, the predictions from the first hypothesis remain confirmed.

In order to give the subject an expectancy of success on Maze A, he was told that his pretest performance was "very good" and "well above average." Logically, there is reason to suspect that subjects who completed a small percentage of the 24 pretest mazes would be less inclined to believe the experimenter's statement than those who completed a larger percentage. In order to test this suspicion, the subjects were grouped by a median split on the pretest (21 subjects in high pretest and 20 subjects in low pretest) and the

<sup>6</sup> Personal communication, June 1971.

data relevant to the experimental hypotheses were reexamined.

The results of this dissection reveal that the first hypothesis regarding the relationship between I-E scale and length of stay is strongly supported in the high pretest sample ( $r = -0.65$ ,  $p < 0.01$ ), but not at all in the low group ( $r = 0.02$ ). However, while manipulation effects seem to be an important factor in the weakening of the relationship between the I-E scale and length of stay, it is still probable that the hypotheses concerning task satisfaction (Hypotheses 3a and 3b) were incorrect regardless of this experimental contamination.

A final analysis involved the multiple correlations between various combinations of the pretest, I-E scale and JDI with length of stay on Maze B. This analysis revealed that the combination of I-E scale and pretest accounted for a significantly greater proportion of variance in length of stay ( $R = 0.56$ , and the population estimate of  $R = 0.54$ ) than either the I-E scale ( $r = -.43$ ) or pretest ( $r = -.38$ ) alone. No other combination of variables resulted in such an increase in prediction. It is apparent that the knowledge of work-related personality variables, as well as ability variables, may result in more accurate predictions of task-related performance. While the results were significant, it must be kept in mind that the analysis was post hoc, and a more controlled study of these relationships should be carried out before more definite conclusions are made.

Although two of the three main hypotheses were supported, the strengthening of the relationships produced by the internal analysis points out the need for more effective procedures. The data indicate that a more convincing expectancy manipulation will have the effect of strengthening the relationship between I-E Control and expectancy change. In addition, any further research done along these lines should have purer measures of ability and original effort incorporated into the design.

#### REFERENCES

ARVEY, R. D., & DUNNETTE, M. D. Task performance as a function of perceived effort-performance and performance-reward contingencies. (Tech. Rep.) Office of Naval Research, 1970.

- ATKINSON, J. W. *An introduction to motivation*. Princeton: Van Nostrand, 1964.
- FEATHER, N. T. Effects of prior success and failure on expectations of success and subsequent performance. *Journal of Personality and Social Psychology*, 1966, 3, 287-298. (a)
- FEATHER, N. T. The study of persistence. In J. W. Atkinson & N. T. Feather (Eds.), *A theory of achievement motivation*. New York: Wiley, 1966. (b)
- GALBRAITH, J., & CUMMINGS, L. L. An empirical investigation of the motivational determinants of task performance: Interactive effects between instrumentality-valence and motivation-ability. *Organizational Behavior and Human Performance*, 1967, 2, 237-257.
- HACKMAN, J. R., & PORTER, L. W. Expectancy theory predictions of work effectiveness. *Organizational Behavior and Human Performance*, 1968, 3, 417-426.
- HERSCH, P. D., & SCHEIBE, K. E. The reliability and validity of internal versus external control as a personality dimension. *Journal of Consulting Psychology*, 1967, 31, 609-613.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. B. *The motivation to work*. New York: Wiley, 1959.
- HILL, J., BASS, A., & ROSEN, H. The prediction of complex organizational behavior: A comparison of decision theory with more traditional techniques. *Organizational Behavior and Human Performance*, 1970, 5, 449-462.
- KATZELL, R. A., BARRET, R. S., & PARKER, T. C. Job satisfaction, job performance, and situational characteristics. *Journal of Applied Psychology*, 1961, 45, 65-72.
- KIESLER, C. A., COLLINS, B. E., & MILLER, N. *Attitude change*. New York: Wiley, 1969.
- KUHLEN, R. G. Needs, perceived need satisfaction opportunities and satisfaction with occupation. *Journal of Applied Psychology*, 1963, 47, 56-64.
- LAWLER, E. A. III, & PORTER, L. W. Antecedent attitudes of effective managerial performance. *Organizational Behavior and Human Performance*, 1967, 2, 122-142.
- LEWIN, K. *The conceptual representation and the measurement of psychological forces*. Durham: Duke University Press, 1938.
- ROTTER, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monograph*, 1966, 80(1, Whole No. 609).
- ROTTER, J. B., SEAMAN, M., & LIVERANT, S. Internal versus external control of reinforcement: A major variable in behavior theory. In N. F. Washburne (Ed.), *Decisions, values and groups*. New York: Macmillan, 1962.
- SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *Measurement of satisfaction in work and retirement: A strategy for the study of attitudes*. Chicago: Rand McNally, 1969.
- VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

(Received October 13, 1971)

## PROFESSIONAL EMPLOYEES' PREFERENCE FOR UPWARD MOBILITY

DOROTHY N. HARLOW<sup>1</sup>

*University of South Florida*

This research attempts to explain why some engineers are more interested than others in upward mobility. The hypotheses were based on Robert Presthus' accommodation theory. Fifty-four graduate engineers completed questionnaires measuring job satisfaction (JS), ambiguity tolerance (AT), and promotional preference (PP). For the total sample, PP and JS were positively correlated ( $p < .05$ ), supporting the theory. For those above the median on JS, AT and PP were also positively correlated ( $p < .01$ ), contrary to the theory. However, AT and PP were negatively correlated ( $p < .01$ ) for 33 engineering managers. Although data from professional employees did not support the theory regarding AT, it is possible that the individual's career stage also should be considered.

This research, a field study of graduate engineers, was conducted to develop knowledge about the types of professional employees who might want to move into management.

Engineers, (herein termed "professionals") after a long, specialized educational process, have above-average status and income. Many other industrial employees, to gain comparable advantages, would have to go into management. For an engineer to move into management, however, could be tantamount to exchanging one attractive organizational role for another.

### ACCOMMODATION THEORY

This research attempts an empirical test of Robert Presthus' theory (1962) of organizational mobility. Although Presthus did not use the label, it is referred to here as accommodation theory. Unfortunately, the entire theory is not reduced to propositions that can be tested in organizations; rather, the ideas are presented in descriptive form.

Presthus' discussion pertains to upward mobility in general and is couched in the familiar local versus cosmopolitan terms. In this field study, the focus was narrowed for the subjects to a specific firm and one organi-

zational level. The design specified the reference point of promotional preference as only one level above that currently occupied by the subject. Organizational space was intentionally limited to only that position in management with which the subjects were most familiar and through which they must move in most firms to gain whatever level the individual might eventually desire to reach. To this extent, the design, while operational and realistic for the subjects, was at variance with accommodation theory.

The theory predicts the influence of job satisfaction and ambiguity tolerance on promotional preference or upward mobility. It is hypothesized that (a) job satisfaction is positively related to preference for promotion, and (b) promotional preference is negatively related to tolerance for ambiguity for individuals who have high job satisfaction.

Presthus, charging that large bureaucratic organizations manipulate people and force them to accept socialization and internalization of the firm's values, contends that individuals must adapt to this environment in order to survive. This adaptation can take three forms: upward-mobiles—desiring promotion very much; indifferents—caring not at all; and ambivalents—both attracted and repelled by such a possibility.

### *Upward-Mobiles*

The upward-mobile, Presthus proposes, is typically a "local," with interests and aspira-

<sup>1</sup> The data reported here is part of that collected in conjunction with the author's dissertation for the University of Kansas, 1970.

Requests for reprints should be sent to Dorothy N. Harlow, Department of Management, College of Business Administration, University of South Florida, Tampa, Florida 33620.

tions tied to the organization. He would have high job satisfaction because he has received the rewards (not specified) of the organization, and he would have a low tolerance for ambiguity. This intolerance, Presthus believes, is an expression of the upward-mobile's deference for authority.

### *Indifferents*

The blue-collar indifferent rejects advancement because he refuses to accept the responsibility involved. His main interests are his off-job activities, and he works primarily for income to support these interests. The indifferent maintains his identity by withdrawing psychologically from the organization.

Specifically related to the hypotheses, Presthus writes:

The indifferent's rejection of status and prestige values often insure a felicitous accommodation. Since job satisfaction (as Presthus defines it) is a product of the relation between aspirations and achievement, he is often the *most satisfied* of organization men [p. 218].

Additionally, the indifferent does not respect authority because he is the product of a lower-class family where, according to Presthus, such respect typically is not taught. The indifferent, then, would have a high tolerance for ambiguity, the opposite of that predicted for the upward-mobile.

Of particular interest to this research, Presthus also held that the professional can be indifferent to advancement, apparently as a result of abuse by the institution or organizational environment. Presthus used college professors as an example of this type of person. The professor, in his alienation, turns his loyalty to his profession, which can be practiced in any institution. He is thus a "cosmopolitan."

### *Ambivalents*

The third ideal type was not involved in the hypotheses because, according to Presthus, the ambivalent is irresolute about promotion.

## METHOD

### *Operational Definitions*

A "professional" is defined as a person whose task assignment is based on a relatively specialized technical undergraduate background. Subjects in this study were individuals with at least an undergraduate degree in engineering and no stipulated minimum work experience, who were below the first level of supervision, and whose current task assignment was in some section of engineering.

"Promotion" is defined as movement upward to the level immediately above one's present organizational assignment with responsibility for subordinates doing work like that currently performed by the subjects. Advancement either in pay or to a higher "working engineer" classification is not defined as a promotion.

The operational definition for "promotional preference" was the response to one questionnaire item: "Assume that in the very near future your immediate superior's job (or its equivalent) became available. If it were offered to you do you think you would accept it?" Response possibilities ranged from "definitely yes" (scored 7), to a "definitely not" (scored 1). No neutral or midpoint was provided; nonresponses were scored as 4. Those engineers having a relatively high or low preference for promotion could be identified by comparing their individual rank score on this item with the median rank position of the total sample.

"Management," as used here, pertained to those organizational line positions responsible, among other things, for evaluating and rewarding subordinates. Individuals in staff positions assigned to coordinate work assignments, but not in control of sanctions, were not considered to be in management.

Accommodation theory holds that job satisfaction is the result of the relationship between expectations for (aspirations) and actual receipt of rewards (achievement). This variable resembles a discrepancy score. While the data of this study did not contain a measure of expectations concerning rewards, the response options did provide subjects an opportunity to express their degree of satisfaction with rewards received. The job satisfaction index used to test both hypotheses was the sum of response categories of the 34 unique items which had been most important for the four factors (supervision, intrinsic, social service, and financial-advancement) of Sedlacek's (1966) professional-managerial subsample.

Accommodation theory specifies ambiguity tolerance as a personality dimension important in determining an individual's attitude toward advancement. This variable was included in the second hypothesis. Total score on the scale developed by Budner (1962) was the measure of "ambiguity tolerance." To make the data compatible with Presthus' theory, the scoring was reversed to provide a high value for tolerance of ambiguity.

One of Budner's statements provides the impres-

sion that his instrument matches the description by Presthus of the upward-mobile.

Empirically, the ambiguity scale was shown to correlate with conventionality . . . The scale also correlated positively with authoritarianism and expressed attitudes of idealization of and submission to parents . . . [pp. 49-50].

### *Collection and Analysis of Data*

This study was conducted in a midwestern city. Eight firms employing both graduate engineers and engineering managers were invited to participate. Of the 137 questionnaires provided the firms, 98 (75%) were returned. Questionnaires were completed voluntarily during regular work hours. The sample contains 54 male graduate engineers below the first level of supervision. Thirty-four respondents were managers. Eleven other questionnaires had to be eliminated primarily because those replying were not graduate engineers.

Mainly because of the sample size, nonparametric statistics seemed justified, and the Kendall Tau was used for data analyses.

Of the 54 engineers, 59% were age 35 or under, and 41% either had accumulated some graduate course work toward or had completed an advanced degree.

The possibility of promotion into management in general was attractive to these employees. The median response category was "probably yes," and 50% of all engineers checked the "definitely yes" category.

### RESULTS

A correlation (Kendall tau) between job satisfaction and promotional preference of .30 ( $p < .05$ ) supports the first hypothesis. That is, those people whose scores on job satisfaction ranked high relative to the other subjects were also the engineers who ranked relatively high on promotional preference.

To test the second hypothesis, job satisfaction scores were ranked in descending order for the total sample. A Kendall tau was computed for promotional preference and ambiguity tolerance for only those subjects whose scores were in the median rank position or above on job satisfaction. This data division was made to test that portion of accommodation theory that postulated that the most satisfied employees are the indifferents and the upward-mobiles, and that these two ideal types can be identified or further separated on the basis of their tolerance-intolerance of ambiguity. The Kendall tau was .308 ( $Z = 2.250, p < .01$ ) for the relationship between promotional preference and am-

biguity tolerance. Although the correlation was significant, the results are contrary to the accommodation theory, which predicts a negative relationship between these two variables. For this sample, those engineers who were most tolerant of ambiguous situations were also those who most preferred upward mobility.

Interestingly, age correlated negatively with promotional preference ( $\tau = -0.232, Z = -2.481, p < 0.003$ ), but seniority was not significantly related.

A criterion validity check of the single item measuring the dependent variable, promotional preference, included responses to Schaffer's (1953) need scale for dominance. Kendall tau correlations were calculated between responses to the 11 items of the dominance scale in Schaffer's Occupational Attitude Survey (which had been collected for another part of this study) and the one promotional preference item. Four dominance items were identified that not only correlated significantly and positively with each of the other three but also with the promotional preference item as well ( $Z$  values of 2.778, 2.829, 1.858, and 1.892, all significant beyond the .04 level). The promotional preference score and total dominance score also were positively and significantly related ( $Z = 2.208, p < .02$  level).

### ENGINEERING MANAGERS

Although the primary interest of the study was promotional preference of professionals, data also were available from 33 graduate engineers who were managers of engineers. These men were from the same firms as the engineers but, because they represented several levels of management, their responses were not included in the tests of the hypotheses. Of the managers above the median on job satisfaction, ambiguity tolerance and promotional preference were, as Presthus predicted, negatively and significantly related ( $Z = -2.328, p < .01$ ).

This relationship for the managers might result from either identification with authority or in response to some personality defense mechanism. It could be that intolerance of ambiguity is related more closely to authoritarianism through the enculturation process

in the work place than by socialization in the culture as a whole, as Presthus proposes. Perhaps in suggesting that upward-mobiles could be distinguished from the indifferents on the basis of this variable (ambiguity tolerance), Presthus may have been in error only in terms of the career stage where the variable is important.

It might be interesting, then, and worthwhile, to test the above speculation using a "before- and after-promotion" research design. Such a design might resemble closely the research involving union stewards and shop foremen which provided support for the Katz and Kahn (1966, p. 189) position that the role shapes the attitudes and perceptions of the individual.

Another possible explanation of the results from the manager group is that the selection process of these firms may have tended to nominate for promotion those who were highly intolerant of ambiguity.

#### DISCUSSION

The potential sample size of professional employees was limited in this midwest city because, for the promotional preference item to have any meaning, the work groups to which the subjects were assigned needed to be large enough to justify a supervisor. Other limiting sample size factors were that many graduate engineers in the area of the study were not employees of firms but worked as independent professionals. Others were assigned to sections or departments within the firm where their engineering training was not being utilized. Data not related to the two hypotheses reported here were also collected, and the total questionnaire required approximately 1 hour to complete.

While only one geographical location was represented, there was no reason to suspect that the variables in the study would be responsive to regional cultures. It also is believed that the firms included in the study are representative of the industries which primarily employ professional engineers: chemical, public utilities, consumer products, petroleum refinery construction, municipal government (civil engineers), and aircraft (private and military).

The inclusion of geographically concen-

trated firms may have been an asset rather than a limitation of this study. Charles Hulin (1966), for example, presents some relatively new job satisfaction evidence in which he claims some benefits can be gained by using a concentrated sample. Job satisfaction is often considered a complex variable. To the extent that promotional preference is also a complex variable, the fact that all responses for this study were from one labor market then may have been an advantage. Hulin correlated satisfaction scores with an index frame of reference based on community characteristics. Hulin concludes:

The results of this study would seem to indicate that conceptualization of job satisfaction which does not include recognition of the part played by frames of reference or alternatives available to the worker is going to be inadequate . . . [p. 191].

The importance of Hulin's results for this research is that all of the subjects shared the same community characteristics. The state of the labor market, for example, was essentially the same for all respondents; the engineers were also, at least momentarily, contemporaries in the same midwestern subculture.

#### *Implication for Management*

Three of the findings in this research seem to be of particular relevance to organizations employing professional personnel:

1. Nearly one half of the professional sample had an intense desire for a management position. It is highly unlikely that enough openings could exist in most firms to accommodate all those employees who view upward mobility as an attractive possibility. One suggestion pertaining to a reversal of this trend would be to better acquaint employees with the actual responsibilities and duties of a first-line manager, anticipating that such an increased awareness would reduce the desire for promotion.

2. Age correlated negatively with promotional preference; seniority was not significantly related. These two results when considered together suggest that a young professional's age is more important than is his length of service in explaining his desire for advancement.

3. An organization should be able to use

the results of this study if it were known, for example, whether currently successful managers of professionals scored high or low on ambiguity tolerance. Those subordinates with high job satisfaction would probably be the most likely to accept a promotion if it were offered, and the candidate's ambiguity tolerance score might be used as a success predictive measure.

### *Implications for Accommodation Theory*

This study indicates Presthus was correct in predicting that professional employees who have a high level of job satisfaction are also the ones who want to advance. It perhaps should be emphasized that the positive correlation in the results of this study between job satisfaction and promotional preference are independent measures: job satisfaction items referred to satisfaction with attributes of one's current position; promotional preference pertained to an organizational role presently occupied by the engineer's own supervisor or manager.

A serious doubt was generated regarding Presthus' speculation about tolerance for ambiguity and its relation to preference for promotion, at least to the first level of management. Exactly opposite from Presthus' position, the "working engineers" in this sample who were most tolerant of ambiguous situations were also those who had the highest preference for promotion.

Presthus' description of the ambiguity tolerance-intolerance variable and the impact he believes it will have on behavior actually follow more closely both the definition and the research results of *authoritarianism* than they resemble *ambiguity intolerance*. Authoritarianism and intolerance of ambiguity are often positively related in research (e.g., Budner, 1962, pp. 41-42). Budner contends that rather than a simple variable, "the nature of the concept . . . (ambiguity tolerance) posits a complex, multidimensional construct [p. 35]." All the Budner scale items tapped at least one of four postulated indicators of perceived threat and at least one of

"three types of ambiguous situations: novelty, complexity, and insolubility [p. 32]."

Budner used peer evaluations from high school students in one validity study for his scale. Three of the items asked for nominations of those in the class who most preferred the status quo (intolerance of ambiguity) in tackling problems while two offered nomination possibilities for those who most like (a) the new and unfamiliar and (b) complex and challenging situations. "The correlation between the peer-rating index and the ambiguity scale was .34 . . . [p. 37]." The last two items were scored as measures of ambiguity tolerance and would most closely represent the movement into management for the engineers of this study. Budner's results are also consistent with other research (Bogen, 1961; Rydell, 1966) which demonstrates that tolerance for ambiguity often accompanies a willingness to change one's opinion and to tolerate new experiences. Only additional research can answer the issue proposed here: Accommodation theory perhaps should be revised to include references to authoritarianism rather than intolerance for ambiguity.

### REFERENCES

- BOGEN, I., JR. Some operational definitions of intolerance of ambiguity and their relationship to adaptation and anxiety. (Doctoral dissertation, University of Denver), Ann Arbor, Mich.: University Microfilms, 1961, No. 61-6594.
- BUDNER, S. Intolerance of ambiguity as a personality variable. *Journal of Personality*, 1962, 30, 29-50.
- HULIN, C. L. Effects of community characteristics on measures of job satisfaction. *Journal of Applied Psychology*, 1966, 50, 185-92.
- KATZ, D., & KAHN, R. *The social psychology of organizations*. New York: Wiley, 1966.
- PRESTHUS, R. *The organizational society: An analysis and a theory*. New York: Knopf, 1962.
- RYDELL, S. T. Tolerance of ambiguity and semantic differential ratings. *Psychological Reports*, 1966, 19, 1303-1312.
- SCHAFFER, R. Job satisfaction as related to need satisfaction in work. *Psychological Monographs*, 1953, 67(14, Whole No. 364).
- SEDLACEK, W. E. An empirical description of available theory and research on job satisfaction. Paper presented at the meeting of the Midwestern Psychological Association, Kansas City, May 1966.

(Received September 27, 1971)

## EFFECT OF HOME ENVIRONMENT TOBACCO SMOKE ON FAMILY HEALTH

PAUL CAMERON<sup>1</sup>

*University of Louisville*

DONALD ROBERTSON

*California State College, Long Beach*

This study replicated and extended earlier research that indicated a greater prevalence of respiratory illness among children subjected to tobacco smoke in the home environment. A random phone sample of 2,626 households in Detroit, Long Beach, and Pasadena, yielded evidence that (a) children subjected to tobacco smoke in the home environment have a greater prevalence of acute illness when compared to children in smoke-free environments, (b) adult nonsmokers subjected to tobacco smoke in the home environment may have a greater prevalence of acute illness than adult nonsmokers who reside in a smoke-free environment, and (c) respiratory illness rates may be related to air pollution rates in metropolitan areas.

This article reports a further study of the relationship of tobacco smoke in the home to the prevalence of illness in children. It represents an attempt to replicate the earlier finding that there is a greater prevalence of acute illness in children subjected to tobacco smoke in the home environment than for those children not subjected to tobacco smoke in the home environment (Cameron, 1967; Cameron, Kostin, Zaks, Wolfe, Tighe, Oselett, Stocker, & Winton, 1969). Only two areas of the country, Denver and Detroit, have been sampled in previous studies; this study samples Detroit, Long Beach, and Pasadena.

### METHOD

As it seems tentatively established that there is a reasonable degree of commonality between random phone sampling and random area sampling in the establishing of acute illness rates (Cameron et al., 1969), we randomly drew 2,150 phone numbers from the Detroit metropolitan, 363 numbers from the Long Beach, and 161 from the Pasadena phone books (a larger number from both Long Beach and Pasadena had been planned but various difficulties aborted the samples). Each number was called and households with children under the age of 17 in residence were sampled via the report of an adult member. As we were mainly interested in testing the hypothesis that children residing in home environments with tobacco smoke present suffer a greater prevalence of acute illness, the cutoff age of 16 was employed for comparability with Public Health Service (PHS) surveys. If a business or an inappropriate household (i.e., no children under age 17 in residence) was called, the next phone number

down the column was substituted. All appropriate families who refused an interview were called back to a maximum of 7 times at which point they were considered a refusal. In Detroit, there were 20 refusals (<1%), in Long Beach, there were 8 refusals (2.2%), and in Pasadena, there were 20 (8%). Any drawn phone number that did not answer was called back on 3 different days to establish the number as "dead" for sampling purposes. Interviewers were college student volunteers who had been trained in the administration of the questionnaire. Twenty-five percent of the data was verified by the recalling and readministration of parts of the questionnaire by junior investigators.<sup>2</sup> All sampling took place from November 3 to November 28, 1968.

After an introduction in which the interviewer identified himself as a representative of the National Health Survey, he asked questions concerning demographic variables, family health, smoking habits, ventilation and pollution. A major difference between our and the PHS acute illness questionnaire is that our procedure required the respondents to report on the health of the family for the past 7 days, instead of the past 14 days as required by PHS.

Coding was always done in the same order as the questions were asked so that the coder did not know whether he was coding a person subjected to smoke or not before he coded them ill or not ill. Further, only about 1% of the responses required any interpretive coding—the categories used by the PHS correspond with those used by the general populace (i.e., if the interviewee characterized an illness as a "cold," it was coded as a "cold"; if the illness was said to be the "flu," and vomiting occurred a great deal, it was coded as "influenza with digestive manifestations"—all interpretive codings are in the "other" illness categories).

<sup>1</sup> Requests for reprints should be sent to Paul Cameron, Department of Psychology, University of Louisville, Louisville, Kentucky 40208.

<sup>2</sup> We wish to thank Mark Berkley, Laura Briscoe, Joe Stolar, Alan Sugarman, David Wattenberg, Bob Rosenbaum, Bernie Webberman, and Christine Mueller for doing the tremendous amount of verification, recalling, and coding without financial reward.

## HYPOTHESIS TESTING EXPLANATION

Before reporting our results, we should probably note that we did not test a non-directional hypothesis with the chi-square statistic. Rather, all chi-squares were derived by testing the specific hypothesis, "Is the respiratory illness prevalence for children subjected to tobacco smoke in the home environment greater than that for children not subjected to tobacco smoke in the home environment?" We felt that the consistency of our previously reported results on this question made testing the null hypothesis of no difference between the two samples a procedure wasteful of information. Therefore, we used the empirically uncovered rate of illness for non-smoke-subjected children in each age grouping to generate the expected prevalence for the smoke-subjected children (e.g., if at a given age category, the non-smoke children had a prevalence of 5%, the chi-square was performed between the expected prevalence at 5% versus the actually obtained figure with  $df = 1$ ). Since the null hypothesis can be rejected because sample "a" is either too great or too small relative to sample "b", the greater efficiency of a specific hypothesis, which eliminates essentially half of the non-directional hypothesis, is obvious.

## RESULTS

Turning first to the question of whether children subjected to tobacco smoke in the environment have a greater prevalence of respiratory illness than children not subjected, our results rather firmly suggest an affirmative answer. Table 1 summarizes the acute illness rate from each of the three locations. All differences that reach statistical significance are in the affirmative direction. Since the Long Beach and Pasadena samples are rather small to detect reliably a difference for an effect on illness prevalences, the spottiness of results is to be expected. Of additional interest, the prevalence of respiratory illness for children who themselves smoke in each location is greater than that of children who are merely subjected to "second-hand" smoke (7.8% for Detroit, 15.4% for Long Beach, and 28.6% for Pasadena smokers).

Chronic illness prevalences for children subjected to smoke were much the same as those

TABLE 1  
THE HEALTH OF CHILDREN SUBJECTED TO TOBACCO SMOKE IN THE HOME VERSUS THE HEALTH OF CHILDREN NOT SO SUBJECTED

Age	No. subjected	No. not subjected	$\chi^2$
Detroit			
10-16	1,506 (77 smokers)	785	
Respiratory illness	104 (6.9%)	38 (4.8%)	13.43***
Acute illness excluding injuries	143 (9.5%)	49 (6.2%)	
6-9	798	356	
Respiratory illness	83 (10.4%)	30 (8.4%)	3.78**
Acute illness excluding injuries	105 (13.2%)	41 (11.5%)	
0-5	933	423	
Respiratory illness	159 (17.1%)	53 (12.6%)	32.0****
Acute illness excluding injuries	195 (20.9%)	64 (15.1%)	
Long Beach			
10-16	189 (13 smokers)	109	
Respiratory illness	12 (6.3%)	10 (9.2%)	-2.88*
Acute illness excluding injuries	27 (14.3%)	11 (10.1%)	
6-9	110	58	
Respiratory illness	7 (6.4%)	1 (1.7%)	9.03***
Acute illness excluding injuries	8 (7.3%)	4 (6.9%)	
0-5	152	87	
Respiratory illness	20 (13.2%)	9 (10.3%)	-1.03 (ns)
Acute illness excluding injuries	22 (14.5%)	13 (14.9%)	
Pasadena			
10-16	78 (7 smokers)	65	
Respiratory illness	15 (19.3%)	5 (7.7%)	6.96***
Any type of illness	18 (12.3%)	8 (5.7%)	
6-9	51	32	
Respiratory illness	5 (9.8%)	2 (6.3%)	.07 (ns)
Any type of illness	7 (13.7%)	6 (18.7%)	
0-5	41	38	
Respiratory illness	4 (9.8%)	6 (15.8%)	-.75 (ns)
Any type of illness	5 (12.2%)	6 (15.8%)	

\*  $p < .10$ .

\*\*  $p < .05$ .

\*\*\*  $p < .01$ .

\*\*\*\*  $p < .001$ .

of children not subjected (in Detroit, for example, for 16-year-olds and under, the prevalences were 1.8% and 1.9% with the lower prevalence favoring the smoke-subjected children).

TABLE 2  
PREVALENCE OF RESPIRATORY ILLNESS FOR  
CHILDREN AGED 16 OR UNDER BY  
DIAGNOSTIC CATEGORY

Kind of Respiratory Illness	Subjected (N = 3,857)	Not subjected (N = 1,953)
Common cold	296 (7.7%)	117 (6.0%)
Other acute upper respiratory illness	14 (0.4%)	3 (0.2%)
Influenza with digestive manifestations	14 (0.4%)	4 (0.2%)
Other influenza	68 (1.8%)	23 (1.2%)
Pneumonia	7 (0.2%)	2 (0.1%)
Bronchitis	7 (0.2%)	3 (0.2%)
Other acute respiratory conditions	13 (0.3%)	2 (0.1%)

Table 2 combines the child data for the three cities by kind of respiratory illness. For each category of respiratory illness, a lower prevalence was obtained for children not subjected to tobacco smoke in the environment (sign test places the probability of 7 out of 7 comparisons favoring the non-smoke-exposed at less than .01).

The question of whether an adult's health is adversely affected by residing with a smoker while not smoking himself gets some answer from the data presented in Table 3. For each city, the health of nonsmokers over the age of 17 residing in a household where one or more of the other members smoked is compared with nonsmokers residing in a smoke-free household. Unfortunately, Detroit was the only location with a large enough sample to enable a reasonable test of the question, and the statistically significant difference favors nonsmokers not subjected to tobacco smoke in the home. As with the children, chronic illness rates were essentially the same for both groups (in Detroit, for instance, the rates were 2.8% and 3.0%).

It will be noted that the percentage of adults subjected to others' household smoke is between 22% and 23% for each location.

Respiratory illness rates of smokers approximated those of nonsmokers. In Detroit 7.1% of smokers had a respiratory illness; in Long Beach the figure was 7.4%; while in Pasadena it was 4.8%.

Table 4 compares adult male smokers and nonsmokers residing in Detroit on each of

the other items of the questionnaire. The median yearly income and average number in household for Detroit families as reported by the United States Census Bureau is recorded in the last column of Table 4. Clearly our phone-drawn sample was comparable to the census sample. There were essentially no differences in the two populations along any of the dimensions (the mean ages were statistically different and nonsmokers averaged about \$200 more income, but neither difference seems large enough to account for the health differences). The "pollution problem" question turned out to be poorly cast and many mentioned water pollution, noise pollution, and the like. Therefore, the equivalent percentages for smokers and nonsmokers suggest equivalent confusion and little else (like Long Beach and Pasadena comparisons similarly yielded no differences).

The hint of an association between amount of tobacco smoke exposure and the prevalence of acute illness for children subjected to smoke uncovered in the last study (Cameron et al., 1969) did not reappear in the present. The biserial correlation between the amount of smoke that sick children under 10 were subjected to versus the amount of smoke

TABLE 3  
ACUTE ILLNESS PREVALENCE FOR ADULT NONSMOKER  
RESIDING WITH TOBACCO SMOKE PRESENT Versus  
ADULT NONSMOKERS RESIDING IN TOBACCO-  
SMOKE-FREE HOMES

Illness	Subjected (N = 1,179)	Not subjected (N = 1,312)	$\chi^2$
Detroit			
Respiratory Acute excluding injuries	80 (6.8%)	75 (5.8%)	
	116 (9.9%)	104 (7.9%)	5.05*
Long Beach			
Respiratory Acute excluding injuries	(N = 159) 10 (6.3%)	(N = 152) 12 (4.8%)	
	15 (9.5%)	15 (6.0%)	2.68 (ns)
Pasadena			
Respiratory Acute excluding injuries	(N = 74) 3 (4.0%)	(N = 143) 12 (8.4%)	
	5 (6.8%)	17 (11.9%)	-1.13 (ns)

\* p < .02.

that smokers' well children were subjected to was quite low (.05) and not statistically significant.

The question of a possible differential bias between self-report and other-report of illness prevalence in large samples of families is partially confronted in Table 5. It should be noted that a given adult falls into either the self-or-other-reported side of Table 5; that is, we do not have here a direct comparison of self- versus other-report of illness prevalence for adults from the same families, but rather a comparison of self- versus other-report prevalence rates for adults from different families. Nonetheless, the obvious lack of a difference between the prevalencies reported in the two arrays argues against the notion that self-reported illness prevalences in a random sample of families will differ from other-reported illness prevalences in another random sample of families from the same population. When the data were further broken down into smokers' reports on other smokers' health versus nonsmokers' reports on other smokers' health the same lack of difference appeared. For reports of children's illnesses the same lack of differences was demonstrated for the Detroit, Long Beach and Pasadena data in separate analyses.

About 75% of the interviews were conducted with the woman of the house, while

TABLE 4  
DEMOGRAPHIC-ENVIRONMENTAL DIFFERENCES  
BETWEEN ADULT DETROIT MALE  
SMOKERS AND NONSMOKERS

Demographic factor	Smokers (N = 1,295)	Nonsmokers (N = 1,043)	Detroit population*
Average ages			
% who regularly take vitamins	45.0	42.9	
% who report below average ventilation	33	31	
% who report a special pollution problem	2.6	2.5	
Median yearly income	14	15	
Average number in household	\$7,500-\$9,999	\$7,500-\$9,999	\$8,800

\* Computed from the 1970 *Statistical Abstract of the United States*, United States Census Bureau, 1970.

most of the remainder were conducted with the man of the household. A third of our adult females and 55% of our adult males smoked as compared with 33% and 51% for the United States population of adults (Ahmed & Gleeson, 1970). Thus, about two thirds of our reports were provided by nonsmokers.

Our study also provided a limited test of the notion that respiratory illness rates should be related to the quality of environmental air. The PHS nationwide, and the Air Pollution Control District in Los Angeles, have published estimates of air pollution that would seem to rank the locations involved as follows: Pasadena, most; Detroit, next; and

TABLE 5  
PREVALENCE OF SELF-REPORTED ILLNESS AND JUDGMENT-OF-ILLNESS BY ANOTHER

Factor	Smokers	Self-report		Report by another		
		Nonsmokers not subjected to second-hand smoke in home	Nonsmokers subjected to second-hand smoke in home	Smokers	Nonsmokers not subjected to second-hand smoke in home	Nonsmokers subjected to second-hand smoke in home
Acute illness						
Total sample	422	259	203	728	375	322
With illness	29	17	25	48	30	27
With illness (%)	6.9%	6.6%	12.3%	6.6%	8.0%	8.4%
Chronic illness						
With illness	22	12	4	32	9	11
With illness (%)	5.2%	4.6%	2.0%	4.4%	2.4%	3.4%

Long Beach, least polluted.<sup>3</sup> If we regard our age-smoke condition groupings (children aged 0-5, 6-9, 10-16 subjected and not subjected to tobacco smoke in the home environment, children who smoke, adults who smoke, adults who do not smoke but are subjected to the same, and adults who do not smoke and are not subjected to smoke) as 10 independent tests of the notion and expect the respiratory illness rate in each group to run lowest in Long Beach and highest in Pasadena, we have 30 predictions and 19 "hits." If we apply Jonckheer's (1954) test to the data (e.g., the first 11 terms of the trinomial expansion  $(1\ 2\ 2\ 1)^{10}/6^{10}$  or  $8,478,468/60,446,176$ ), we arrive at a probability of .14. Thus, the data fall in the right direction but fail of statistical significance.

### DISCUSSION

The presence of tobacco smoke in the home environment seems to be generally associated with a greater prevalence of respiratory illness in children. The effect has now been found in three rather dissimilar metropolitan areas with varying climates, altitudes, and types of air pollution. The possibility that summer might find the effect diminished is strengthened by the relative weakness of the difference between smoke-subjected and non-smoke-subjected children in the Los Angeles area samples. Even though the time period was the same, in November, Detroit was rather cold and not conducive to outdoor play—the opposite of the climatic conditions in

southern California. Because of the pleasant weather, the tobacco smoke in the home was probably less frequently encountered by resident children; thus, both categories of children probably shared outdoor air more frequently in California. It should be noted that most of the children of both samples were not ill at the time of interview. Second-hand tobacco smoke appears to be a significant, but not an all-determining independent variable.

As all our reported research to date has been by phone surveys in metropolitan areas and of the cross-sectional-associational design, it would seem appropriate at this time to mention a piece of longitudinal research done by a graduate student under the direction of the senior author in a rural area of Michigan. Hermann (1968) followed the school absences of the 102 first and fifth graders for the first seven weeks of the fall, 1968, school term in the Hillside Elementary School. She found that the median number of absences ascribed to illness for children subjected to tobacco smoke in the environment ( $n = 74$ ) was higher than for children in smoke-free environments; further, while the median number of half-days absent for children from one-smoker-present families was 0, the corresponding figure for children with two or more smokers in residence ( $n = 37$ ) was 2.

Although non-smoking adults subjected to smoke displayed a statistically greater prevalence of acute illness, it is by no means certain that the finding is a function of the smoke per se. It is possible that the smoke affects their children's health, then the adults "catch" the illness from their children. We will need large samples of smokers with and without children to test this possibility (thus, if we find greater illness among childless non-smokers subjected to smoke, we will have essentially eliminated the latter possibility).

We did not directly confront a possible psychological difference between smokers and nonsmokers that could have generated our results—smokers may be more apt to regard their children as ill at a given intensity of symptoms. That is, smokers may be generally more health-conscious either in general or in regard to their children. Three lines of evidence suggest that this interpretation of our results is not very attractive. First, when the

<sup>3</sup> The Air Pollution Control District of Los Angeles, in a personal communication, reported median single day highs of ozone for each month of the year. The median reading for the West San Gabriel Valley (Pasadena is included here) was .39 while for the South Coastal area (Long Beach) the figure was .17. The United States Department of Health, Education, and Welfare Public Health Service in its August 4, 1967 press release estimated the relative air pollution of the Long Beach-Los Angeles area at 393.5 vs. a 370.0 reading for Detroit. If we assume rough comparability of the two indices (i.e., ozone vs. the PHS conglomerate of various kinds of air pollution), and take at face value that Pasadena is approximately twice as polluted as Long Beach, we would estimate that Pasadena to have a PHS index of approximately 494 and Long Beach an index rating of approximately 246 in which case the Detroit rating would fall in between.

health data for smoke-subjected children were split into smoker-reported versus nonsmoker-reported prevalences, no statistically significant differences emerged. Though the same group of families' health status was not re-indexed using nonsmokers' and smokers' reportage (and the possibility of a reportage-bias still exists), the possibility would seem to be considerably diminished by our finding. Secondly, vitamins have been advertised extensively and are widely believed to be "health-insurance" by the general populace. The essential equivalence of vitamin usage for smoking and non-smoking males (Table 4) (and the same general equivalency obtained for their wives and children) suggests a no-greater health concern among smokers. Lastly, anyone who smokes today must find ways to rationalize or discount mounting scientific opinion that judges his habit health hazardous. While it cannot be maintained that anyone who smokes is *ipso facto* less health conscious, it certainly seems possible that smokers would be somewhat less, rather than more, health conscious. There are other possible differences between the smoke-subjected and non-smoke-subjected families that might have generated some or all of the differences. Among these might be reduced discretionary income (an adult smoker usually spends between \$100 and \$200/year to maintain his habit) that might otherwise go for superior food products or greater household cleanliness (assuming that either affects health), or greater safety consciousness.

It is likely that many physicians reading this account are puzzled at the lack of a significantly higher rate of illness among adult smokers. We would remind them that sickness is a psychosocial event with no necessary physical parameters. It is undoubtedly true that smokers cough more, that their lung functioning is reduced, etc.; yet, such physical phenomena do not constitute illness—unless the person involved and/or his interactants judge him ill. If a person gets used to coughing at a given rate, it makes small difference that most people do not cough that frequently—he is not ill in his own eyes. And

if his family and friends are also used to such a rate for him, he is likewise not sick to them. True, his physiologic functioning may be below average, but there is not necessarily a relationship between illness and physiologic functioning. Thus, if we tested the physiological state of the adult smokers in our sample, we would almost certainly find them below average on many counts; but they and their families are *used* to such a bodily state, and they are simply *not ill* any more frequently. The reason smoking children are most frequently reported ill or report themselves ill is that neither they nor their parents are yet used to their symptoms—predictably both will become used to them and no longer judge the person ill more frequently. Illness, after all, is something only *persons* can have—bodies can deteriorate, machines wear down, but *only people can be sick*.

It would seem profitable to pursue the idea that an association exists between physical health and the quality of environmental air. We are pursuing the possibility that the health differences between smokers' and non-smokers' children will lessen in the summer season.

#### REFERENCES

- AHMED, P. I., & GLEESON, G. A. *Changes in cigarette smoking habits between 1955 and 1966*. Washington, D. C.: Public Health Service, 1970.
- CAMERON, P. The presence of pets and smoking as correlates of perceived disease. *Journal of Allergy*, 1967, 40, 12-15.
- CAMERON, P., KOSTIN, J. S., ZAKS, J. M., WOLFE, J. H., TIGHE, G., OSELETT, B., STOCKER, R., & WINTON, J. The health of smokers' and non-smokers' children. *Journal of Allergy*, 1969, 43, 336-341.
- HERMANN, M. M. S. The differences between the health of children residing in a smoky environment and the health of children residing in a non-smoky environment in the Hope-Edenville, Michigan area. Unpublished manuscript, Western Michigan Univ., 1968.
- JONCKHEERE, A. R. A test of significance for the relation between  $m$  ranks and  $k$  ranked categories. *British Journal of Statistical Psychology*, 1954, 7, 93-100.
- OSSORIO, P. G. *Persons*. Los Angeles: Linguistic Research Institute, 1966.

(Received September 7, 1971)

# THE LIFE HISTORY QUESTIONNAIRE AS A PREDICTOR OF PERFORMANCE IN NAVY DIVER TRAINING<sup>1</sup>

ROBERT HELMREICH<sup>2</sup> AND ROGER BAKEMAN

*University of Texas*

ROLAND RADLOFF

*National Science Foundation*

A new demographic instrument, the Life History Questionnaire (LHQ) is described. The LHQ elicits demographic data longitudinally providing a question-by-year matrix of responses. Variables derived from the LHQ are used to predict success in Navy diver training. The utility of the LHQ both for prediction and as a research tool is discussed.

One of the most widely accepted truisms in psychology is that the best predictor of future behavior is past behavior. Research evidence supports this contention; for example, the best predictor of college grades is high school grades; previous income predicts success in selling life insurance (Tanofsky, Sheff, & O'Neill, 1969); completion of high school predicts completion of service school and Navy enlistment (Plag & Goffman, 1966). Our own previous research has also convinced us of the value of such information. In a study of Aquanaut performance during the Navy's Project SEALAB II, life history items were most successful in predicting performance, especially in contrast with personality and interest inventory data (Radloff & Helmreich, 1968).

Theoreticians have argued the potential power of life history information (see Guthrie, 1944, for an especially compelling argument). More recently it has been asserted

that biographical information is the "best single predictor of future behavior where the predicted behavior is of a total or complex nature [Henry, 1966]."

The demonstrated utility of life history information appears to have resulted in more concern with employing biographic variables in applied situations such as counseling and personnel selection than with exploring the conceptual properties underlying such information. Owens (1968), in particular, has focused on this issue. Supporting the notion that life history information has underlying conceptual consistency is the fact that factor-analytic studies have shown significant structural relationships among biographic items (e.g., Baehr & Williams, 1967; Morrison, Owens, Glennon, & Albright, 1962; Schmuckler, 1966; Thomson & Owens, 1964).

The impetus for the development of the Life History Questionnaire was a large-scale field investigation of the behavior of Aquanauts during Project TEKTITE 2 (Helmreich, 1971). Our goal was to understand and explain differences among TEKTITE Aquanauts in their ability to work effectively underwater, to get along with fellow teammates, and to adjust generally to a stressful, isolated and confining environment. Since we were attempting to predict complex real-life behavior, it followed that the best predictive information would be a total record of prior experiences. We looked for and failed to find extant measuring instruments which would yield such information in a consistent longitudinal format.

<sup>1</sup> This research was funded by the Organizational Effectiveness Research Programs, Psychological Sciences Division, Office of Naval Research under Contract No. N00014-67A-0126-0001, Contract Authority Identification No. NR171-804, Robert Helmreich, Principal Investigator. The study was conducted while Roland Radloff was Research Psychologist, Naval Medical Research Institute, Bethesda, Maryland. We wish to express our particular gratitude to Lieutenant Thomas Berghage, Chief Warrant Officer William Dool, Ensign William Weeks, Lieutenant Robert Biersner, Lieutenant Commander T. Murray, and Lieutenant John Whitaker who provided invaluable assistance in data collection.

<sup>2</sup> Requests for reprints should be sent to Robert Helmreich, Department of Psychology, University of Texas, Mezes Hall 211, Austin, Texas 78712.

### The Life History Questionnaire

The Life History Questionnaire (LHQ) was conceived and designed to assess experience and behavior during the first 19 years of a person's life. Its intent is to elicit comprehensive information by covering such areas as place of residence; size of hometown; frequency of moves; type and size of residence; size and composition of family; quality of food and clothing; father's and mother's employment, education and occupation; comparative height and weight; health; type and size of school; school performance; participation in athletic and other activities; religious participation; frequency of going out at night and dating; fights with peers; clashes with authority; parental praise, criticism, physical affection, and punishment; work and financial independence (see Table 1).

Two major influences guiding the selection of areas to be covered were: *A Catalogue of Life History Items* (Owens, Glennon, & Albright, 1966) and a factor analytic study of the dimensions of personal background data (Baehr & Williams, 1967).

Questions in the LHQ emphasize the occurrence of events rather than attitudes and feelings. For example, "In what size community did you live?" rather than "In what size city would you prefer to live?"; or "How often did your parents punish you?" rather than "How strict did you feel your parents were?"

Qualitative responses can also dilute factual information, as noted by Owens, Glennon, and Albright (1962). Qualitative responses result when response categories such as "never, seldom, frequently, often or very often" are used. The problem is, of course, that one man's "frequently" is another man's "seldom." In the LHQ, wherever possible, responses are coded in numerical frequencies such as: once per year, once per month, once per week, daily, etc.

An essential feature of the LHQ is the provision for year-by-year responses. Twelve questions are answered 19 times, once for each year. The other 20 questions ask for responses only for appropriate years, as in questions on dating, school attendance, and school performance. The use of multiple responses per-

TABLE 1  
LIFE HISTORY QUESTIONNAIRE ITEMS

Item	Age range (by year)
Multiple response	
1. Geographical residence	0-18
2. Hometown size	0-18
3. Distance of home from larger population centers	0-18
4. Type of residence	0-18
5. Condition and status of residence	0-18
6. Family size and composition	0-18
7. Clothing quality	0-18
8. Food—quantity and quality	0-18
9. Father's employment	0-18
10. Mother's employment	0-18
11. Height	0-18
12. Weight	0-18
13. Health	0-18
14. Education—type of school	5-18
15. Education—size of school	5-18
16. Education—academic performance	5-18
17. Athletic achievement and awards	5-18
18. Intellectual achievement and awards	5-18
19. Other awards and honors	5-18
20. Religious activities	5-18
21. Going out at night	9-18
22. Dating	12-18
23. Fights with peers	5-18
24. Clashes with authority	5-18
25. Financial independence	5-18
26. Work—school year	5-18
27. Work—summer months	5-18
28. Parental praise	5-18
29. Parental physical affection	5-18
30. Parental verbal criticism	5-18
31. Parental physical punishment	5-18
32. Community homogeneity and personal similarity	0-18
Single response	
1. Father's occupation	
2. Mother's occupation	
3. Father's education	
4. Mother's education	
5. Subject's education	
6. Other languages spoken	
7. Height	
8. Weight	
9. Birth month and year	
10. Marital status	
11. Sex	

mits measurement of several important aspects of life history, including: number of changes, direction of changes, rate of develop-

ment, and age at the occurrence of an event. A few examples may illustrate the importance of such information. Later behavior may be influenced as much by improvements or declines in school performance as it is by average performance; as much by rate at which financial independence is achieved as it is by the fact of its achievement; and as much by the age at which parents were divorced or died as it is by the fact of divorce or death. Influences deriving from such factors as the number and direction of changes, rate of development, and age at occurrence of events cannot be known unless a matrix of data is available. Questions answered year-by-year seem to be the most sensitive method of obtaining this information. Although the fallibility of human memory in recalling detailed quantitative information about early experience may weaken results, subjects report little difficulty in retrieving the information and preliminary studies of test-retest reliability show high consistency.

The nature of questions on the LHQ may be illustrated by the question concerning health.<sup>3</sup> The instructions for the question begin "How healthy or unhealthy have you been? For each year of age, indicate the number of days you have been unable to take part in regular activities because of ill health by use of the appropriate number from the categories defined below. Unable to take part in regular activities means being in a hospital; staying home from school or work; staying home on weekends, holidays, or evenings when you might normally have been out of doors, visiting friends, going somewhere for entertainment or recreation, doing errands or similar activities." This is followed by additional information concerning response categories. The response categories used are:

1. Zero days of restricted activity due to ill health.
2. One to 6 days restricted activity due to ill health.
3. Seven to 14 days, 1 to 2 weeks, restricted activities due to ill health.

<sup>3</sup> A revised version of the questionnaire has been published: Radluff, R. & Helmreich, R. The Life History Questionnaire. *JSAS Catalog of Selected Documents*, 1972, 2, 13.

4. Fifteen to 30 days, more than two weeks, up to 1 month restricted activities due to ill health.

5. Thirty-one to 60 days, 1 to 2 months restricted activities due to ill health.

6. Sixty-one to 120 days, more than 2 and up to 4 months, restricted activities due to ill health.

7. One hundred twenty-one to 140 days, more than 4 and up to 8 months, restricted activities due to ill health.

8. Two hundred forty-one or more days, more than 8 months and up to the full year, restricted activities due to ill health.

9. Don't remember.

The first applications of LHQ derived predictors to behavioral criteria were highly successful and have been reported elsewhere for Aquanauts (Helmreich, 1971) and Navy divers (Radloff, 1971). In the present article, we will present an application of the data available from the LHQ to prediction of completion and relative standing in two demanding military schools training Navy divers, second class.

## METHOD

Subjects were 115 male enlisted men in the United States Navy who composed five classes in training to be Divers, second class.<sup>4</sup> This school population is composed of volunteers and presents basic instruction in SCUBA diving for the Navy. All subjects were given the LHQ at the beginning of the training course. At the end of the 10 week course, criterion information was collected for each trainee. The criteria were completion or noncompletion of the course and class rank for those successfully completing training.

### Scoring and Coding LHQ Data

The LHQ is answered on a machine readable answer form from which responses are automatically transcribed onto punch cards producing a matrix of yearly responses. In addition to the response matrix, several background questions such as father's education, subjects current weight, etc., are answered only once. These data are then processed by program LIJAHN (Bakeman, 1972). This program permits the investigator to extract from combinations of

<sup>4</sup> The first and third classes were from the United States Navy School of Diving and Salvage, Washington, D.C.; the second, fourth, and fifth were from the United States Navy Diving School, Harbor Clearance Unit 2, Norfolk, Virginia.

TABLE 2  
PRODUCT-MOMENT CORRELATIONS OF PREDICTORS WITH CRITERIA (VALIDATION SAMPLE  $N = 52$ )

Factor	1	2	3	4	5	6	7	8	9	10	11	12
1. Parental affection <sup>a</sup>	-.27											
2. Educational performance <sup>a</sup>	.00	.16										
3. Health <sup>a</sup>	.24	-.31	.02									
4. Athletic honors <sup>a</sup>	.09	.00	.05	.21								
5. Weight <sup>a</sup>	.07	.04	.02	.20	.95							
6. Weight/height												
7. Difference in parents' education	-.36	.22	-.03	-.15	-.11	-.05						
8. Birth order	-.21	.00	.09	-.04	.13	.11	-.02					
9. Social status index	.13	-.33	-.17	.09	.02	.04	.02	-.07				
10. Home family index	.00	.01	.02	.02	.09	.11	.10	-.15	.27			
11. Pass-Fail criterion	.40	-.20	.20	.36	.17	.20	-.16	-.25	.16	.25		
12. Performance criterion	.37	-.20	.20	.44	-.07	-.10	-.12	-.16	.03	.10	.84	

<sup>a</sup> Based on mean of responses for ages 13-17 inclusive.

raw data of the LHQ values of a priori variables for each subject.

To do this, conceptual variables must first be defined. This is done by constructing a table where each entry or line in the table describes a different conceptual variable. For each conceptual variable, the user indicates: (a) the statistic to be computed; these include mean, median, mode, change scores, and trend scores; (b) the LHQ questions to be used; and (c) within that question, the years to be considered in the analysis.

Given the data available from the LHQ, an almost limitless number of conceptual variables could be formed; in practice, only a few would be. Research hypotheses and previous experience will typically suggest appropriate variables. Here, we have allowed our prior experience with divers to guide conceptual variable definition. Since the present study is intended primarily as an exploration of the use of the LHQ, we have deliberately defined only a limited number of variables.<sup>5</sup>

## RESULTS

The first two classes studied were assigned to the validation sample ( $N = 52$ ). The next three classes ( $N = 63$ ) were assigned to the cross-validation sample. 61% (32) of the validation sample completed training successfully. Fifty-seven percent (36) of the cross-validation population completed the course.

Two criteria were formed for the samples. The first, a pass-fail indicator, was coded with 0 = failure (disenrollment from the course because of inability to meet classroom or diving standards), 1 = pass (successful

<sup>5</sup> An archive of LHQ's with criterion data is being accumulated from a number of diverse populations. Factor analyses of longitudinal responses will be undertaken when the subject  $N$  per population exceeds 500.

fulfillment of all course requirements and certification as a second class diver). A broader, 4-point performance criterion was formed with: 1 = noncompletion; 2 = completion in the bottom one-third of the class; 3 = completion in the middle one-third; and 4 = completion in the top one-third.

Four predictors were computed as the mean of yearly responses between the ages of 13 and 17 inclusive. These were: (a) Parental Affection (mean number of occasions when parents expressed physical affection); (b) Educational Performance (relative secondary school class rank); (c) Health (coded as mean number of days restricted due to illness or accident); and (d) Athletic Honors (mean number of recognitions for athletic endeavor).

Two variables designed to reflect socioeconomic status were formed from LHQ items. The first (called Social Status) was the sum of father's educational level, mean quality of food served in the home, and mean quality of clothing provided for the subject. The second variable (called Home-Family Index) was computed by subtracting the mean number of persons living in the nuclear family from the mean number of rooms in the family domicile.

Four additional variables were based on single response items on the LHQ. These were: subject weight; the weight-height ratio (weight divided by height in inches); difference in parental education (father's educational level minus mother's educational level); and birth order (firstborn vs. later born).

The correlations of predictors with the criteria for the validation sample are shown in Table 2. The multiple regression analyses were conducted using the SPSS regression program (Nie, Bent, & Hull, 1970). For the pass-fail criterion, all 10 predictors met the inclusion requirement and yielded a multiple correlation of .60. The cross-validity of the predictive equation was .58 with a standard error of estimate of .44. Both samples were highly comparable on all predictive variables; no differences approached significance.

The multiple correlation with the performance criterion was .61 in the validation sample. The multiple correlation in the cross-validation sample was .59 with a standard error of estimate of 1.25.

#### DISCUSSION

Correlations with the criteria provide some indication of characteristics associated with success in a rigorous diving course. The high positive correlation between receiving physical affection and success in military training is interesting not only because it demonstrates a strong relationship between family atmosphere and the criteria but also because it seems to support the contention that the objective format of the LHQ can provide quantitative information about rather subjective experiences. The correlations between educational performance and the criteria are in the expected direction, higher class rank is associated with course completion and performance. IQ scores (in the form of scores on the Navy's General Classification Test) were available for some subjects ( $N = 38$ ). The correlation between the IQ measure and the pass-fail criterion was nonsignificantly negative (-.11). This implies that the LHQ question concerning school performance is more related to achievement motivation than to academic intelligence.

The relationship between health and criteria, although nonsignificant, is in a counter-intuitive direction and replicates a finding obtained with Scientist-Aquanauts during Project TEKTITE 2 (Helmreich, 1971). This is a tendency for successful performers to have experienced considerable restriction because of illness or accident. Another aspect of this relationship between health and per-

formance illustrates one of the major capabilities of the LHQ. In a detailed analysis of the relationship between the health variable at different age periods and performance among TEKTITE Aquanauts, it was found that the statistical effect was produced by a strong relationship between illness during elementary school years (6-12) and the criterion. Among Aquanauts, the relationship was much weaker both in early childhood and during teenage years. In examination of data for the Navy sample, the same effect is noted. The correlation between the health variable and the pass-fail criterion for years 6-12 was .32 while the correlation for ages 0-5 years was .16. One implication is that restriction during early school years with subsequent recovery may lead to an emphasis on physical achievement. In any event, the LHQ data facilitate the exploration of such questions concerning the relative importance of experiences at different ages.

Birth order was significantly related to the pass-fail criterion with primogeniture associated with success (also replicating the effect found among Aquanauts; Helmreich, 1971). The LHQ provides extensive data on ordinal position and sibling structure. However, because of the limited sample size, only a dichotomous predictor was formed.

The Athletic Honors variable correlates positively with the criterion. This quantitative measure of athletic accomplishment relates strongly to the physical task of diving. The two variables relating physique to completion and performance show moderate relationships in the not-surprising direction that heavier and stockier (higher weight-height ratio) divers are somewhat more likely to pass. The relationship to the performance criterion is much weaker, probably indicating a threshold effect. That is, a stocky diver is likely to pass, but beyond that, the extent of his stockiness does not predict how well he will do.

The two socioeconomic predictors were also more strongly related to the pass-fail criterion than to the performance measure. Higher socioeconomic status is associated positively with the pass-fail criterion, but only weakly with the performance measure. This distinc-

tion between variables that predict attainment of an acceptable level of performance and those that predict variations in levels of excellence seems both a practical and a theoretically fascinating one. Clearly, it calls for further investigation.

The ability of variables derived from the matrix of information available from the Life History Questionnaire to predict performance in diver training suggests that the LHQ may be a highly useful tool. Studies are currently underway relating the LHQ to performance in other settings. However, the most important feature of the instrument appears to be the fact that it provides sufficient longitudinal data to enable detailed investigation of the relationships among a variety of life settings and experiences and to relate these to subsequent behavior. With a large data base, many developmental and social hypotheses can be systematically explored.

#### REFERENCES

- BAEHR, M., & WILLIAMS, G. Underlying dimensions of personality background data and their relationship to occupational classification. *Journal of Applied Psychology*, 1967, 51, 481-490.
- BAKEMAN, R. *Processing the Life History Questionnaire* (ONR Tech. Rep. No. 19) Austin: University of Texas, 1972.
- GUTHRIE, E. R. Personality in terms of associative learning. In J. McV. Hunt (Ed.), *Personality and the behavior disorders*. New York: Ronald Press, 1944.
- HELMREICH, R. *The TEKTITE 2 Human Behavior Program*. In Miller, J., Vanderwalker, J., & Waller, R. (Eds.), *TEKTITE 2: Scientists in the sea*. Washington: Government Printing Office, 1971.
- HENRY, E. Conference on the use of biographical data in psychology. *American Psychologist*, 1966, 21, 247-249.
- MORRISON, R. F., OWENS, W. A., GLENNON, J. R., & ALBRIGHT, L. E. Factored life history antecedents of industrial research performance. *Journal of Applied Psychology*, 1962, 46, 281-284.
- NIE, N. H., BENT, D. H., & HULL, C. H. *SPSS: Statistical package for the social sciences*. New York: McGraw-Hill, 1970.
- OWENS, W. A. Toward one discipline of scientific psychology. *American Psychologist*, 1968, 23, 782-785.
- OWENS, W. A., GLENNON, J. R., & ALBRIGHT, L. E. Retest consistency and the writing of life history items—a first step. *Journal of Applied Psychology*, 1962, 46, 329-332.
- OWENS, W. A., GLENNON, J. R., & ALBRIGHT, L. E. *A catalog of life history items*. For Scientific Affairs Committee, American Psychological Association, Division 14. Reproduced by Creativity Research Institute of the Richardson Foundation, June, 1966.
- PLAC, J., & GOFFMAN, J. The prediction of four-year military effectiveness from the characteristics of naval recruits. *Military Medicine*, 1966, 131, 729-735.
- RADLOFF, R., & HELMREICH, R. *Groups under stress: Psychological research in SEALAB II*. New York: Appleton-Century-Crofts, 1968.
- RADLOFF, R. *Life history and success in diving school*. (Research Rep.) Bethesda, Md.: Naval Medical Research Institute, February 1971.
- SCHMUCKLER, E. *Age differences in biographical inventories, a factor analytic study*. Greensboro, N.C.: The Creativity Research Institute, Richardson Foundation, 1966.
- TANOFSKY, R., SHEPP, R., & O'NEILL, P. Pattern analysis of biographical predictors of success as an insurance salesman. *Journal of Applied Psychology*, 1969, 53, 136-139.
- THOMSON, R. W., & OWENS, W. A. A factorial study of the life history correlates of engineering interests. *American Psychologist*, 1964, 19, 478. (Abstract)

(Received October 1, 1971)

# COLOR VERSUS NUMERIC CODING IN A KEEPING-TRACK TASK: PERFORMANCE UNDER VARYING LOAD CONDITIONS<sup>1</sup>

JACELYN WEDELL<sup>2</sup> AND DAVID G. ALDEN<sup>3</sup>

*Systems and Research Center, Honeywell, Inc., St. Paul, Minnesota*

The effectiveness of a color code versus a numeric code was investigated in a modified keeping-track task. The task studied was that of the air traffic controller. Altitude state was the coded variable. It was hypothesized that color coding would be superior to numeric coding, particularly with a greater number of total items displayed. It was further hypothesized that color would be relatively more efficacious with a greater number of items in the interrogated state. Neither of these hypotheses were supported. Based on an error-type analysis, it was concluded that color can aid in retaining information concerning category size and item spatial location; identity information was quickly lost. The design implications of these findings were discussed.

During the past decade the task of an air traffic controller has become increasingly more difficult as aircraft speed, altitude capabilities, and number of aircraft have increased. The controller's primary information source is a radar display that contains position, altitude, and aircraft identification information. The position information is shown directly by the location of the radar return or "blip" with respect to the center of the display and the surrounding area. Altitude and identification information is contained on the scope face or on small plastic tabs placed next to the blip by the controller. All three pieces of information are critical to the successful completion of the controller's task of keeping track of all aircraft assigned to him and aircraft which enter his sector by mistake.

While he is controlling these aircraft, position will change and altitude may change. Position changes are obvious to the controller because the blip moves. However, altitude changes cannot be observed directly, but must be read from accompanying information. A reduction in workload could be achieved if a more direct means of coding aircraft to match assigned altitudes was found.

<sup>1</sup> Research was conducted at Hamline University, St. Paul, Minnesota.

<sup>2</sup> Now at the Department of Psychology, University of Oregon.

<sup>3</sup> Requests for reprints may be sent to the second author at Honeywell, Inc., Systems and Research Center, 2345 Walnut Street, St. Paul, Minnesota 55113.

The present study was designed to examine the effectiveness of a color code versus a numeric code in a keeping-track task. It would seem that introduction of an effective chunking and coding scheme could reduce the difficulty of the task and increase the accuracy of the controller. It is hypothesized that color coding might enable the operator to quickly encode and update information.

Yntema and Mueser (1960, 1962) and Yntema (1963) have extensively studied the short-term memory task in which one is required to keep track of the present states of a number of variables whose values are changed at random intervals. One of their findings is that an operator performs at his best when there are few variables with many possible states rather than many variables with few possible states. Results of a study by Alden, Wedell, and Kanarick (1971) have shown that a redundant color code does not yield a significant improvement in performance, as compared to performance with symbol or color codes alone. The subjects in the redundantly coded groups reported that they often ignored the redundant code and made use of it primarily when spatially encoding information (e.g., when two adjacent channels contained the same color). The authors suggested that increasing complexity through increasing the number of channels subject has to keep track of may result in a change in a keeping-track strategy. The subject would be required to organize groups of items on some

dimension, i.e., "chunk" (Miller, 1956). The redundant color code would provide an additional basis on which to chunk. The results suggested that a primary color code would also be of value for purposes of organizing information into categories.

Clark (1969) found that color was of greater value in reporting by category than are numbers. Her results supported the hypothesis that color information can be immediately coded neurally on stimulus presentation. Information having verbal content, numerically or alphabetically coded, must be analyzed before encoding while, color-coded information can be immediately encoded by a sensory storage system.

It was hypothesized that color coding would be superior to numeric coding of category altitude information particularly as item load increases, for it permits the partitioning of the display into discrete categories. The number of aircraft at the interrogated altitude state was also varied. Either one, two, or three aircraft were in the state subject was questioned about. It was hypothesized, in addition, that as the number of items per interrogated category increased, color would be the more efficient memory aid.

## METHOD

### Subjects

The subjects were 36 male students from Hamline University, pretested for normal color vision by the Farnsworth D-15 test for color defectiveness. Subjects were assigned randomly to one of the six treatment combinations according to the experimental design.

### Experimental Design

The experimental design was a Myers' (1966) three-factor nested design with two between and one within subject variables. The two between factors were coding condition (A), with two levels (numeric and color coding) and aircraft load (B), with three levels (6, 8, or 10 aircraft). The numeric-six load group kept track of six aircraft whose altitude states were coded by the numbers one through six. The numeric-eight load group kept track of eight aircraft, etc. Similarly the color load groups kept track of 6, 8, or 10 aircraft whose altitude states were color coded. The within-group factor, interrogated state load (C) had three levels (1, 2, and 3). These levels refer to the number of aircraft in an altitude state at the time that state was interrogated. Six independent groups of six subjects each were run. The number of items in those states not

being interrogated was balanced so as not to exceed three.

### Apparatus

Stimulus materials were presented in the form of 35 mm. color slides back-projected on a ground glass screen.

The slides were produced by photographing a 6 × 6 inch square matrix on which Munsell color chips were arranged. The matrix background was Munsell N7 (grey). The following six Munsell colors were selected for relative discriminability: 5R 4/14 (red); 2.5PB 5/8 (blue); 2.5YR 6/14 (orange); 10GY 5/8 (green); 5Y 7/8 (yellow); and 2.5P 5/8 (purple) based on the estimates of an independent sample of observers ( $n = 3$ ). Pressure sensitive Letraset instant letters, Grotesque 216, 14 pt., were applied to the color chips. These chips were used in the color coding group. Additional Munsell N7 chips, coded with pressure sensitive letters of the same font and numbers, DOL 252, 14 pt. were used for the numeric coding group. A total of 120 slides were prepared for each of six groups: 4 instruction slides, 36 practice slides, 8 warm-up slides, and 72 test slides. A 35 mm. Honeywell Pentex camera mounted on a Burke and James Tri-Dimensional copy enlarging and reducing camera equipped with a Gaze 300 mm. Red Dot Artar lens and Kodak Wratten No. 81A filter was used to photograph the stimulus materials. Kodak 35 mm. high speed Ektachrome film was used. The camera length was focal f16 with an exposure of 1/250 sec.

The slides were mounted in Gepe double glass slide binders and placed in slide trays. A modified Kodak Carousel 35 mm. slide projector model number 800 equipped with a 3 inch Kodak Ektanar lens and automatic control was used to project the slides onto a Polacoat rear-projection glass screen, type L540-120G 1/4. The screen was mounted on a wooden table so that the subject was seated comfortably at the table 22 inches away.

Illumination from the screen was sufficient to write by. The projected matrix was 10.25 inches square. Each letter or number was .283 × .233 inch when projected.

Response sheets coded for the subject number, practice or test run, and trial number were prepared for each subject and inserted in a ring notebook. The sheets were white,  $8\frac{1}{2} \times 11$  inches with a  $6 \times 6$  inch matrix printed on them.

### Procedure

After pretesting for color vision defectiveness, the subject was seated at the table with the response notebook before him, and the instructions read to him. With the instructions, four slides were shown: (a) a blank  $6 \times 6$  inch matrix, (b) an illustration of all six coding states (colors or numerals); (c) a sample update (information) slide and (d) a sample interrogation (question) slide.

After all of subject's questions were answered, the room lights were dimmed and the practice slides

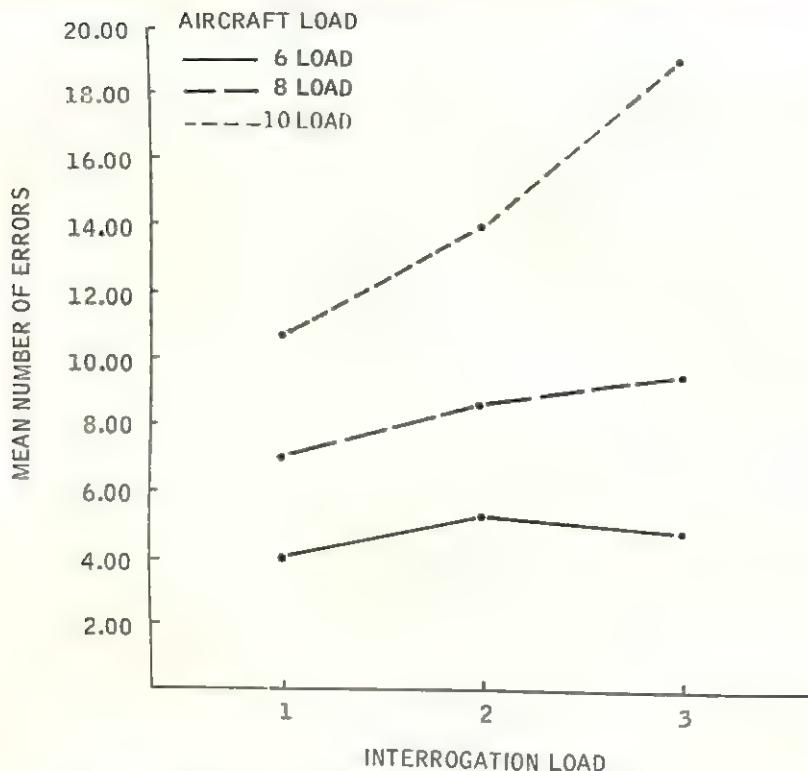


FIG. 1. Mean errors by aircraft load as a function of interrogation load, summed over coding condition.

shown. Eighteen update and 18 interrogation slides were included in the practice set.

Each set (whether practice or test) consisted of alternating update and interrogation slides. The subject was informed that two changes occurred each time a new update slide was shown: His task was to detect these changes and incorporate them into his grouping of aircraft. The interrogation slides consisted of either a color or a number on a neutral background. When such a slide appeared, subject was to respond with the capital letters identifying those aircraft currently in that altitude state and record them in their proper positions on the response sheet. He was to turn the sheet over before the next update slide appeared.

All slides were shown for 15 seconds each, with a 1 second black-out between successive slides while the projector advanced. A new set of response sheets was inserted in the notebook during the break between the sets of practice and test slides.

The first series began with four initial update (information) slides and four interrogation slides which were not scored. The remaining 36 update trials were balanced for equal presentation of each aircraft and interrogation load state so that two observations of each possible combination were displayed to each subject. The sequence was also controlled so that no more than three aircraft were in the same altitude state at any one time. No state or altitude was interrogated more than twice in a row.

## RESULTS

### Scoring

The number of incorrect responses per interrogation was used as the measure of performance. Therefore, response matrices were scored according to type of error made. Errors were categorized according to one of five types: (a) correct aircraft identity, incorrect position (I); (b) incorrect identity, correct position (II); (c) incorrect identity, incorrect position (III); (d) omission (IV); and (e) commission (V).

For the initial analysis, however, error types were combined. A response scored under Error Type I would have been a correct aircraft (capital alphabetic letter) placed improperly on the response sheet. Error Type II implies the reverse: a response in the correct cell of the matrix but incorrectly identified. Error Type III means that a response was given, but neither identity nor position were correctly reported.

Errors were placed in the last two categories only if the number of responses did not

agree with the interrogated load, e.g., if the interrogated load (number of aircraft at an altitude state at the time that altitude state was interrogated) was three and subject made two responses, one error of omission would be recorded. Likewise, if the interrogated load was two, and subject made three responses, one error of commission would be recorded. Thus, a perfect score would be zero; up to three errors were possible for each interrogation.

For the initial analysis of variance error types were combined. Homogeneity of vari-

ance between the six coding aircraft load groups was verified by Hartley's test (Winer, 1962), ( $F_{max} = 3.05$ ,  $df = 6/17$ ,  $p < .05$ ).

The data were then subjected to a  $2 \times 3 \times 3$  analysis of variance with coding condition and aircraft load as between subject variables and interrogation load as the within subject variable (Meyers, 1966). The results yielded significant effects for aircraft load ( $F = 15.19$ ,  $df = 2/30$ ,  $p < .001$ ), with mean errors of 14.25, 25.33, and 43.67 for 6, 8, and 10 loads, respectively, as well as for interrogation load ( $F = 11.45$ ,  $df = 2/60$ ,  $p < .001$ )

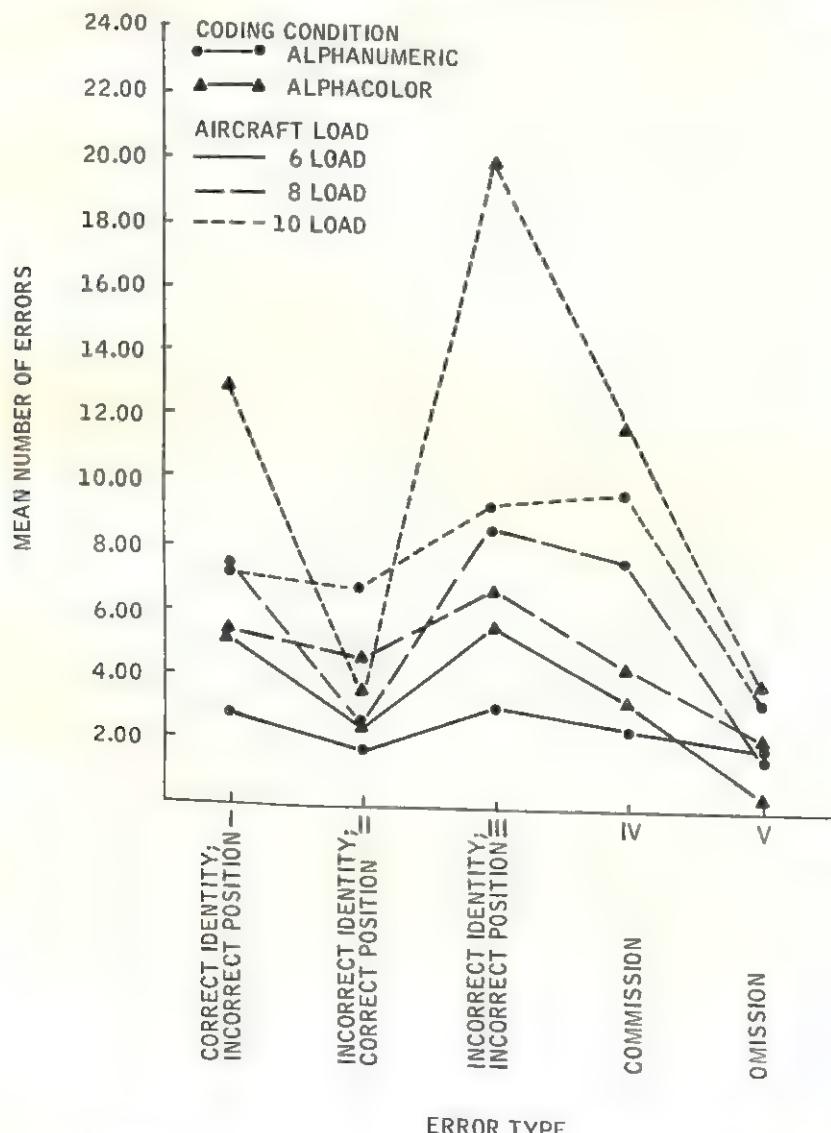


FIG. 2. Mean errors by coding condition and aircraft load as a function of error type, summed over interrogation load.

with means of 7.17, 9.50, and 11.08 for loads of 1, 2, and 3. The Aircraft Load  $\times$  Interrogation Load interaction was also significant ( $F = 4.38$ ,  $df = 4/60$ ,  $p < .05$ ) (see Figure 1). The effect of coding condition was not found to be significant, however, with means of 8.41 and 10.09 for numerics and colors respectively. Coding condition did not interact with any of the loading variables.

The significant Aircraft Load  $\times$  Interrogation Load interaction was investigated by use of a simple main effects test (Winer, 1962). A significant difference was found only for the ten aircraft load condition ( $F = 9.12$ ,  $df = 4/60$ ,  $p < .001$ ). This load condition was apparently so far beyond subject's capacity as to make error scores for the entire aircraft load main effect highly significant.

The data were also analyzed by error type. A Friedman two-way analysis of variance by ranks (Siegel, 1956) was used to compare the six treatment combinations and the five error types. The result was a highly significant difference ( $\chi^2_r(4) = 19.57$ ,  $p < .001$ ). Different types of errors occurred as a result of the conditions subject was performing under, as can be seen in Figure 2. In particular, as the display load increased to 10 items the color coding group could only efficiently keep track of the number of items in the interrogated state.

#### DISCUSSION

Color coding was not found to be superior to numeric coding in either the higher aircraft load or interrogation load states. The results of the error analyses do indicate, however, that color can aid in retaining information concerning the number of items presented. The spatial position of these items was remembered to some degree. However, identity information was quickly lost.

The effect of coding was in the opposite direction to that predicted, the effect being most prominent on the 10 load condition. The significant interaction of error type and coding group indicated that errors were not independent of coding condition. In particular, the greatest number of errors for the color-10 load group was scored under Error Type III: incorrect identity, incorrect position (see Fig-

ure 2). A score in this category indicates that the correct number of responses were given although the spatial location and identity of the aircraft were incorrect. The number of items per state was retained; identity and position information were lost. Further, support of the assumption that identity information is lost before position information is given by the number of errors scored under Type I, again for the color-10 load group. Type I is "correct identity; incorrect position"; Type II is the reverse of this. The color-10 load group was able to retain position information (3.50 mean errors) much easier than identity information (12.66 mean errors). This is in contrast to the number-10 load group that made equal numbers of Type I and II errors.

The results of Clark (1969) indicate that when color category information must be retrieved, color is quickly encoded with location in parallel and thus less subject to decay. When location information must be retrieved, color category information, spatial location information, and identity information are sequentially encoded in respective order. Thus, identity information is most subject to decay, followed by location information, due to the increased time required for encoding.

Similar results are obtained in the present study. Color category information is interpreted here as load information, i.e., how many aircraft are at that altitude? Spatial location information is understood as the actual location of information in the matrix. Identity indicates the identity of the aircraft as given by the capital letter assigned to it. With these definitions established, the sequential processing assumption of Clark (1969) is supported by the present results: Color category information was retained in most instances.

The finding that color was useful for retaining the number of aircraft at specific altitudes is of considerable importance. The subjects were able to group aircraft on the basis of color (altitude state) information. From the standpoint of an air traffic controller's task, this result suggests that color coding might be a practical means of adding a third dimension to a two-dimension display. By

providing color coded altitude information, the controller would know that all aircraft of one color are at the same altitude simply by noting the color of the images. The absolute altitude values in feet would not be readily apparent, but altitude separation would be available in much the same manner as position information. It would seem that by providing this information the total workload of the controller might be reduced especially in regions of crowded air traffic. For example, two aircraft on a collision course are only in danger if they are at the same altitude. This situation requires an immediate response from the controller. At present he must verify the assigned altitude of each aircraft by reading the accompanying information. If the altitude (state) information was color coded, the fact that they are at different altitudes or at the same altitude could be determined directly, reducing the controller's decision time.

Rehearsal of information involves matching of information to its verbal "name" in storage. This verbal "name" is most likely easier to retrieve for numerals than for colors, thus making the matching process more rapid. Apparently, in a keeping-track task, use of a class possessing sequential order (such as numerals) is of greater value than use of a class that can rapidly be chunked on stimulus presentation (such as colors; cf. Monty, Fisher, & Karsh, 1967). Chunking (or categorizing) of information by the use of color does not provide sufficient identity information. As the results of Clark (1969) and the present study indicate, color is much more valuable in carrying information concerning the number of items per category. Identity information is either never encoded or is lost

most rapidly, followed by location and category information.

In summary, the implication of these findings for equipment and display design is that for the transmittal of identity and position information the emphasis should be upon natural language codes readily rehearsable. Color should be used as an aid to search or for an alerting function, or as means to stratify the display, but not as a primary information source.

#### REFERENCES

- ALDEN, D. G., WEDELL, J. R., & KANARICK, A. F. Redundant stimulus coding and keeping-track performance. *Psychonomic Science*, 1971, 22, 201-202.
- CLARK, S. E. Retrieval of color information from perceptual memory. *Journal of Experimental Psychology*, 1969, 82, 263-266.
- MEYERS, J. L. *Fundamentals of experimental design*. Boston: Allyn and Bacon, 1966.
- MILLER, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, 63, 81-97.
- MONTY, R. A., FISHER, D. F., & KARSH, R. Stimulus characteristics and spatial encoding in sequential short-term memory. *Journal of Psychology*, 1967, 65, 109-116.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- WINER, B. T. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.
- YNTEMA, D. B. Keeping track of several things at once. *Human Factors*, 1963, 5, 7-17.
- YNTEMA, D. B., & MUESER, G. E. Remembering the present states of a number of variables. *Journal of Experimental Psychology*, 1960, 60, 18-22.
- YNTEMA, D. B., & MUESER, G. E. Keeping track of variables that have few or many states. *Journal of Experimental Psychology*, 1962, 63, 391-395.

(Received September 13, 1971)

## WRITTEN INFORMATION:

### SOME ALTERNATIVES TO PROSE FOR EXPRESSING THE OUTCOMES OF COMPLEX CONTINGENCIES

PATRICIA WRIGHT<sup>1</sup>

*Medical Research Council, Applied Psychology Unit,  
Cambridge, England*

FRASER REID

*Psychology Department, Brunel University,  
London, England*

Problems were solved using information written either as: (a) bureaucratic style prose, (b) flow chart or algorithm, (c) a list of short sentences, or (d) a two-dimensional table. Prose was always slower to use and more error-prone than other versions, but for nonprose formats there were interactions with problem difficulty. Easier problems resulted in no differential error-rates, although the table was used most rapidly; for harder problems, the algorithm gave fewest errors. Differences in retention strategies appeared when subjects worked from memory. Here performance with prose and short sentences continued to improve over trials, whereas performance with the algorithm and table deteriorated. It is concluded that the optimal format for written information depends on conditions of use.

Written information plays an important part in any technologically advanced environment. Operating instructions, technical manuals and handbooks are indispensable to the smooth running of many appliances, both at work and at home. Nevertheless, written instructions are often presented in ways that are difficult for the reader to follow, understand or remember (Chapanis, 1965; Houghton, 1968; Jones, 1964). This might seem inevitable if the subject matter is itself inherently complex. Frequently such instructions involve conjunctive and disjunctive relations between several events. The following experiment compares the effectiveness of alternative ways of presenting the same items of information about complex contingencies.

There are perhaps two standard formats widely used for dealing with this kind of material. It is either written as prose or acquires a style characteristic of bureaucratic publications. The sentences are long and have numerous embedded qualifying clauses: "If . . . then . . . , unless . . . , in which case . . . , except when . . ." Alternatively the material is presented in tabular form, which may require the simultaneous use of both row and column headings. It is known that tabulation

schemes vary in the ease with which they can be used (Wright & Fox, 1970), but no empirical comparisons have previously been made of the relative difficulties of prose and tables. Therefore these two presentation formats are included in the following experiment.

Wason (1962) pointed out that prose could sometimes be rewritten as a flow chart or logical tree. Here the minimum number of binary decisions required to determine a unique outcome are structured in a logical sequence. Lewis, Horabin, and Gane (1967) referred to such formats as "algorithms," and this is the term which will be used throughout this article. Evidence for the superiority of algorithms over some kinds of prose has been presented by Wason (1968) and Jones (1968).

But clearly sentences do not have to be as long and unwieldy as those characteristic of bureaucratic prose. Simplification by the use of shorter sentences (Flesch, 1945) and improvements such as the introduction of sub-headings (Klare, Schuford, & Nichols, 1958) would make the material easier to understand. Whether such shorter sentences would yield performance comparable to that obtained with the algorithm is examined below.

The user of written information, as distinct from the general reader, is typically looking for the answer to some specific question (e.g., searching through a handbook to

<sup>1</sup> Requests for reprints should be sent to Patricia Wright, Medical Research Council, Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, England.

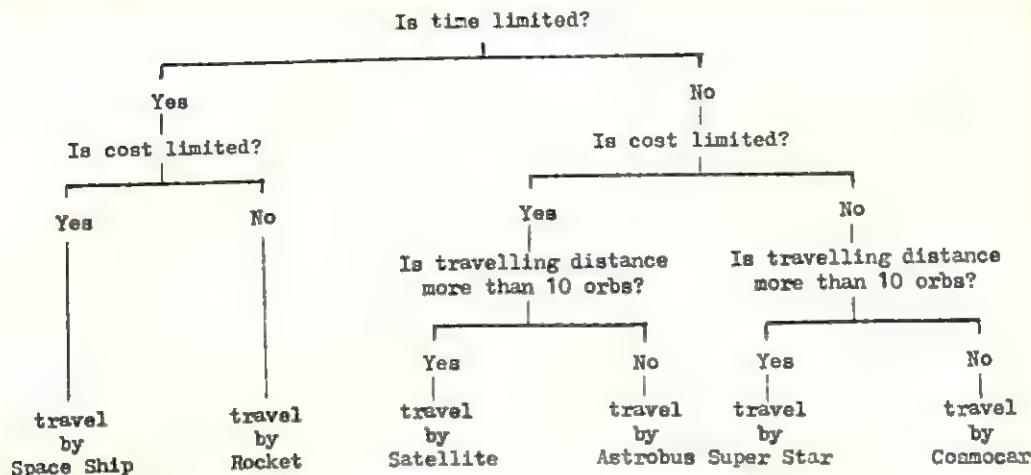


FIG. 1. The algorithm used in the experiment.

discover if the dynamo housing must be dismantled before the ignition element can be renewed). Because characteristics of the problem may interact with the ease of using a particular format, problems at two levels of difficulty were included in the following experiment.

Sometimes written information, of the complexity being considered here, must be committed to memory. It does not follow that information which is easily used when directly available is also easily memorized (Morton, 1967). Consequently both incidental learning and performance after a deliberate effort has been made to remember the material are measured.

Thus the following experiment examines how easily the same basic information can be used when presented either as prose, algorithms, lists of short sentences, or tables, both when the problems are straightforward and when they become more complicated and when the written material is directly available or when it must be used from memory.

#### METHOD

##### Subjects

Sixty-eight adults took part in this experiment, 17 in each of the four experimental groups; 32 were male, 36 were female. Forty-one subjects were paid volunteers from the subject panel of the Applied Psychology Unit, Cambridge; 27 were enlisted naval ratings. In each of the algorithm, table, and prose groups there were 7 enlisted men and 10 volunteers. The short sentences group comprised 6 enlisted men and 11 paid volunteers.

#### Materials

Fictitious material was invented so that subjects had no option but to read the written information to solve each problem; the problems could not be solved from any previous knowledge. The material dealt with the appropriateness of six space vehicles for different kinds of travel.

The algorithm and matrix are shown in Figures 1 and 2, respectively. The Prose passage was intended to approximate the traditional bureaucratic style and read as follows:

When time is limited, travel by Rocket, unless cost is also limited, in which case go by Space Ship. When only cost is limited an Astrobus should be used for journeys of less than 10 orbits, and a Satellite for longer journeys. Cosmocars are recommended, when there are no constraints on time or cost, unless the distance to be travelled exceeds 10 orbits. For journeys longer than 10 orbits, when time and cost are not important, journeys should be made by Super Star.

The list of short sentences was set out as follows:

*Where only time is limited*  
travel by rocket.

*Where only cost is limited*  
travel by satellite if journey more than 10 orbits.  
travel by astrobus if journey less than 10 orbits.

*Where both time and cost are limited*  
travel by space ship.

*Where time and cost are not limited*  
travel by super star if journey more than 10 orbits.  
travel by cosmocar if journey less than 10 orbits.

A series of 36 problems was drawn up. For each problem the subject had to specify the mode of travel appropriate for the situation described on the card. Twelve cards gave the information directly; e.g.,

	If journey less than 10 orbs	If journey more than 10 orbs
Where only time is limited	travel by Rocket	travel by Rocket
Where only cost is limited	travel by Astrobus	travel by Satellite
Where time and cost are not limited	travel by Cosmocar	travel by Super Star
Where both time and cost are limited	travel by Space Ship	travel by Space Ship

FIG. 2. The table used in the experiment.

"more than 10 orbits, cost is limited, time doesn't matter." The remaining cards gave the information implicitly; e.g., "A renowned safe-cracker had just pulled a job on Saturn and wants to get as far away as possible as quickly as possible. He has a bag full of used pound notes." The ordering of distance, cost and time information varied randomly across all problems.

#### Procedure

An independent groups design was used, each subject working with only one of the information formats. Subjects were tested individually and the experimental session was divided into three consecutive sections. During Section 1 the written information was directly available for inspection. In the first half of this section (12 problems), the specifica-

tion of details within problems was straightforward. The remaining problems in Section 1 required more interpretation by subjects to extract the critical details which then had to be related to the written information. In Section 2, the written information was removed without warning and another 12 problems were presented. These problems were similar to those used in the second half of Section 1. Section 3 began with subjects studying the written material for 5 minutes, knowing that further problems had to be solved without consulting this information. Another 12 problems were then presented; again these problems required interpretation by subjects. For all three sections of the experiment, the main dependent variables were errors and latency. In Section 1, the latency corresponded to the time spent reading the written material; in Sections 2 and 3, the latency measured was from the presentation of the problem to its solution.

The information was presented as a photographic negative in the front of a rear lighted box. Subjects could press a button that turned on the lights and started a timer; when the button was released the lights went out and the timer stopped. Thus the time measured corresponded to the time during which the information was visible and, by implication, the time spent reading the information.

In Sections 2 and 3 of the experimental session, the problem cards were presented at the bottom of a box, the top of which contained a half-silvered mirror. Switching on the lights in the box enabled the problem card to be seen, and also started a timer. An answer button pressed by the subject stopped the timer, providing a measure that corresponded to the problem solution time.

TABLE 1  
MEAN PERCENT ERRORS WHEN WRITTEN MATERIAL AVAILABLE<sup>a</sup>

Problem	Prose	Algorithm	Short sentences	Table
Straightforward				
<i>M</i>	34.4 <sup>a</sup>	18.1 <sup>b</sup>	19.1 <sup>b</sup>	14.7 <sup>b</sup>
Difficult	<i>SD</i>	2.1	2.5	2.6
<i>M</i>	41.7 <sup>c</sup>	26.0 <sup>d</sup>	41.7 <sup>c</sup>	35.8 <sup>d</sup>
<i>SD</i>	3.1	2.1	2.7	2.6

<sup>a</sup> Means having common superscripts are not significantly different from each other ( $p > .05$ , two-tailed).

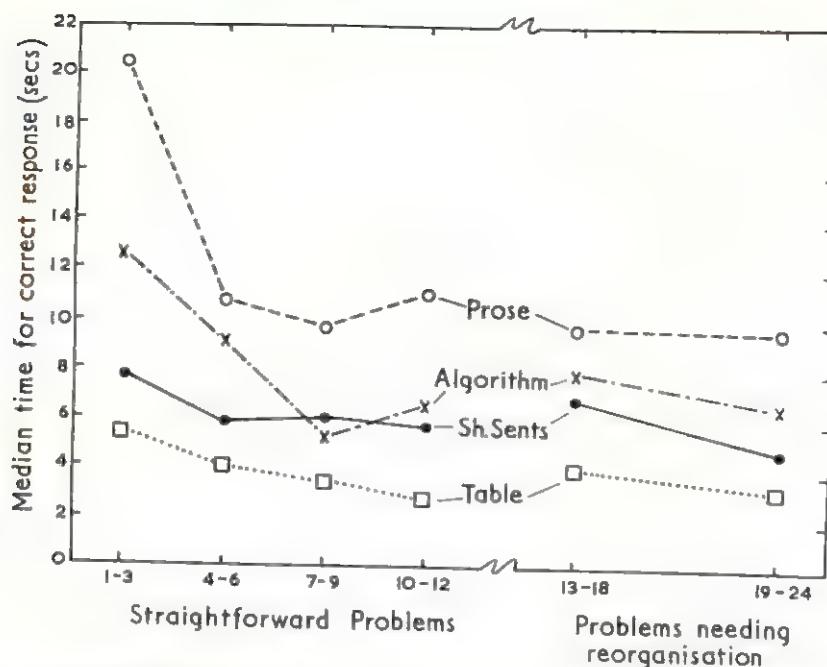


FIG. 3. Median times spent viewing the material prior to correct response (performance on Section I).

## RESULTS

### Section I

Clearly the most important datum is error rate. If information cannot be used accurately there is little value in it being used speedily. Mean error scores for the four types of information are shown in Table 1 together with the standard deviations.

From Table 1 it can be seen that the advantage of the algorithm is greatest when the problems were difficult. Here the algorithm gave significantly fewer errors than the two sentential versions ( $p < .04$ ). For easier problems there were no differential error rates among the nonprose formats, although all were more accurately used than prose. The suggestion in Table 1 that for the straightforward problems the table may have been easier to use than the other versions is confirmed by an analysis of the latency data shown in Figure 3. Correct solutions were obtained more rapidly with the Table than with any other version ( $p < .01$ ).

### Sections II and III

When subjects started working from memory the average error rate across all treat-

ments rose from 36% to 70%. Table 2 shows that very little incidental learning had taken place with any of the materials. This remained true even when the analysis was confined to those subjects who correctly answered at least half the difficult problems on Section 1.

The instruction to memorize the material produced an average error reduction of 11.8% over all treatments. For all materials except the algorithm, the improvement in performance was statistically significant ( $p < .05$ ). Nevertheless in terms of overall performance

TABLE 2  
PERCENT ERRORS WHEN WORKING FROM MEMORY

	Prose	Algorithm	Short sentences	Table
Incidental learning				
<i>M</i>	67.2	68.3	69.9	73.1
<i>SD</i>	1.9	1.8	1.8	1.5
After memorizing				
<i>M</i>	57.8	58.3	53.9	61.3
<i>SD</i>	2.9	3.3	2.9	3.1
Drop in errors due to memorizing	9.4	10.0	16.0	11.8

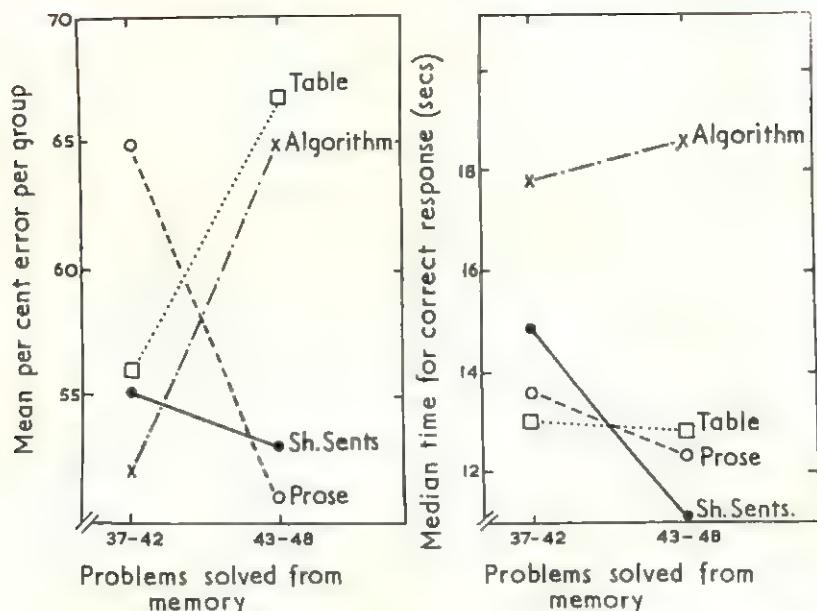


FIG. 4. Errors and correct solution times after the material had been memorized (performance on Section III).

on Section 3, there were no differences between the materials.

However, analysis of overall performance may be somewhat misleading. There were marked differences between materials in the distributions of error scores across trials. The errors made on the first and second halves of Section 3 are shown in Figure 4, where it is evident that errors with the prose and short sentences tended to decrease over trials; whereas the errors of those who had memorized the algorithm and the table tended to increase over trials. The interaction was statistically significant for the comparison of prose and algorithm ( $p = .03$ ) and prose and table ( $p = .02$ ). A similar difference between performance with the sentential and non-sentential materials is shown in the latency data of Figure 4 and may reflect the way in which subjects tried to remember the material.

#### DISCUSSION

The data from Section 1 suggests that when the material was directly available to the user, the formats most easily used were the table and the algorithm. For the easier problems the table was much quicker to use with-

out being any more error prone than the other versions. This contrasts with the findings of Wright (1968) who reported that subjects sometimes had great difficulty in understanding how to use a currency conversion table when it was presented in the form of a two dimensional matrix. Possibly the numerical ability of subjects may have contributed to the difficulty in that particular instance. But there is other evidence that tabulation schemes requiring subjects to co-ordinate two separate pieces of information can cause great difficulties irrespective of whether the tables are numerical or non-numerical (Wright & Fox, 1972). Moreover, a variant of the present table, which was used in a pilot study preceding this experiment, was not well used by some subjects. This pilot table had less redundancy in each cell and more highly condensed row and column headings. The errors included giving as an answer various items of information both from within the cell and from either the row or column heading (e.g., *Rocket if time limited*). Since the time contingency had been specified in the problem it was inappropriate to include a reference to it in the answer.

One conspicuous difference between the pilot table and that used in the present ex-

periment was that with the latter the column headings, row headings and cell contents yielded phrases that when combined approximated an English sentence: e.g., *if journey more than 10 orbs/travel by rocket/where only time is limited*. The corresponding grouping for the pilot table resulted in *travel distance: more than 10 orbs/rocket travel/constraints: only time*. Whether this was the critical difference between the two tables cannot be determined from the present study, but clearly much more needs to be known about two-dimensional matrix formats. The present data suggest that they can sometimes be a very useful way of presenting written information about complex contingencies.

One of the limitations on the usefulness of tables is indicated in the Section 1 performance on the more difficult problems. The algorithm resulted in fewer errors than the other versions when the problem solver had to extract the relevant details from the problem as presented. Therefore, comparison of performance with the algorithm on the two parts of Section 1, suggests that the algorithm has helped the user to structure his problem, rather than assisted him in finding the solution. If this is so, then performance with the other versions might be improved if additional assistance in structuring the problem were provided. One possibility would be to draw the user's attention to the relevant dimensions by listing the three questions: (a) Is travel distance more than 10 orbs? (b) Is cost limited? (c) Is time limited? Such assistance might well enable the advantages of the table (it is quicker to use and requires less space) to be utilized more widely.

That the incidental learning was poor is not surprising when one notes that even after subjects had made a deliberate attempt to learn the subject matter, errors were still above 50% for all materials. Clearly it was not easy to learn.

After the material had been memorized, performance with the algorithm and the table deteriorated over trials. It is possible that subjects relied fairly heavily on visual imagery to encode these formats, and the rising error rate was caused by this visual representation becoming less distinct over suc-

cessive trials. In contrast, performance with the prose and list of short sentences tended to improve over trials. Probably the sentential material was encoded verbally rather than visually, but that would not itself account for the continued improvement. It is possible that the subjective reorganization necessary for memorizing the material (Shiffrin, 1970; Tulving, 1962) did not stop when the material was removed, and this internal re-organization resulted in improved performance.

One final point that must be raised is the question of the generality of the present findings. Earlier in this discussion it was pointed out that the table used was not the only one which met the criterion of presenting these specific items of information, and relatively small changes in the wording of row and column headings could seriously affect performance. Similarly alternative algorithms could be drawn. The one chosen for the experiment capitalized on some of the internal redundancy of the subject matter and, as a consequence, needed only five choice points and six terminal points. If the questions within the algorithm had been in the order of *distance, cost, and time* there would have been seven choice points and eight terminal points (corresponding to the eight cells of the table). It is possible that the greater symmetry of this larger flow chart might have made it easier to remember, or reconstruct from a recollection of the terminal sequence. But even so, if subjects are relying on visual imagery, performance will still deteriorate with time. Moreover there seems no reason for thinking that any variant of the algorithm would result in performance as fast as that observed with the Table on Section 1.

It would seem necessary to conclude that a number of factors combine to determine the optimal way of presenting complex contingencies in the form of written information. Algorithms are useful if the problem to be solved is embedded in miscellaneous information, so that the relevant factors have to be disentangled from the irrelevant. For simpler problems Tables may be preferable to other formats. If information must be used from memory then sentential material may be better remembered. Only one finding recurred

consistently throughout the present study. Alternatives to prose in the traditional bureaucratic style were found to be a considerable improvement, either in time or errors or both.

## REFERENCES

- CHAPANIS, A. Words, words, words. *Human Factors*, 1965, 7, 1-17.
- FLESCH, R. F. More about gobbledegook. *Public Administration Review*, 1945, 5, 240-244.
- HOUGHTON, J. *Round holes not circular orifices*. London: The London and Southern Junior Gas Association, 1968.
- JONES, S. Why can't leaflets be logical? *New Society*, 1964, 102, 16-17.
- JONES, S. *Design of instruction*. London: Her Majesty's Stationery Office 1968.
- KLARE, G. R., SHUFORD, E. H., & NICHOLS, W. H. The relationship of format organisation to learning. *Educational Research Bulletin*, 1958, 37, 39-45.
- LEWIS, B. N., HORABIN, I. S., & GANE, C. P. *Flow charts, logical trees and algorithms for rules and regulations*. London: Her Majesty's Stationery Office, 1967.
- MORTON, J. A singular lack of incidental learning. *Nature*, 1967, 215, 203-204.
- SHIFFRIN, R. M. Memory Search. In D. A. Norman (Ed.), *Models of human memory*, Academic Press, 1970.
- TULVING, E. Subjective organization in free recall of "unrelated" words. *Psychological Review*, 1962, 69, 344-354.
- WASON, P. *Psychological aspects of negation*. London: Communication Research Centre, University College, 1962.
- WASON, P. The drafting of rules. *The New Law Journal*, 1968, 118, 548-549.
- WRIGHT, P. Using tabulated information. *Ergonomics*, 1968, 11, 331-343.
- WRIGHT, P., & FOX, K. Presenting information in tables. *Applied Ergonomics*, 1970, 1, 234-242.
- WRIGHT, P., & FOX, K. Explicit and implicit tabulation formats. *Ergonomics*, 1972, 15, 175-187.

(Received October 18, 1971)

---

Manuscripts Accepted for Publication in the  
*Journal of Applied Psychology*

- Effects of Manipulation of a Performance-Reward Contingency on Behavior in a Simulated Working Condition. Dale O. Jorgenson (Department of Psychology, California State University, 6101 East Seventh Street Long Beach, California 90840), Marvin D. Dunnette, and Robert D. Pritchard.
- Relation of a Test of Attention to Road Accidents. Daniel Kahneman (Oregon Research Institute, 1009 Patterson Street, P.O. Box 3196, Eugene, Oregon 97403), Rachel Ben-Ishai and Michael Lotan.
- Problems of Organizational Control in Microcosm: Group Performance and Group Member Satisfaction as a Function of Differences in Control Structure. Edward L. Levine (Arizona State Personnel Commission, 1831 West Jefferson, Phoenix, Arizona 85007).
- Hiring, Training, and Retaining the Hard-Core Unemployed: A Selected Review. Paul S. Goodman (Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213), Paul Salipante, and Harold Paransky.
- Ego Defensiveness as a Determinant of Reported Differences in Sources of Job Satisfaction and Job Dissatisfaction. Toby D. Wall (Department of Psychology, University of Sheffield, Sheffield S10 2TN, England).
- The Cost of Attaining Personnel Requirements (CAPER) Model: A Method for Evaluating Alternative Recruiting-Selection Strategies. William A. Sands (Department of the Navy, Naval Personnel Research & Development Laboratory, Washington Navy Yard, Washington, D.C. 20309).
- Another Look at Contrast Effects in the Employment Interview. Frank J. Landy (Department of Psychology, Pennsylvania State University, 417 Psychology Building, University Park, Pennsylvania 16802) and Frederick Bates.
- Race, Economic Class, and Perceived Outcomes of Work and Unemployment. Jack M. Feldman (Department of Management, University of Florida, Gainesville, Florida 32601).
- Work Values of White Collar Employees as a Function of Sociological Background. S. D. Saleh (Faculty of Engineering, Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada) and T. Singh.
- Behavior of Temporary Members in Small Groups. Thomas G. Walker (Department of Political Science, Emory University, Atlanta, Georgia 30322).

# MARKET IMAGE AS A FUNCTION OF CONSUMER GROUP AND PRODUCT TYPE: A QUANTITATIVE APPROACH

ALAN POHLMAN AND SAMUEL MUDD<sup>1</sup>

*Smoke Psychological Laboratories, Gettysburg College*

Approximately 30 brands of each of three product types (automobiles, beers, magazines) were rated for "classiness" on a Thurstone Scale. Medians and semi-interquartile range values ( $Q$ ) of the rating distributions were calculated for each brand item. These statistics were found to vary systematically as a function of product type and consumer group. Familiarity was shown to be correlated significantly with some dimensions of the class-rating distributions.

A substantial body of research has accumulated indicating a systematic interaction between preferred product type and/or brand and purchaser variables such as social class (Coleman, 1964), self-concept (Sommers, 1969), and personality (Tucker & Painter, 1969). Interpretations of these findings center around the notion of value placed on product or brand ownership by the purchaser. The purchased item is conceptualized as having two kinds of value for the owner, one for its concrete functional utility and the other for its utility as a prestige symbol. According to this conceptualization, functional value is that which is conventionally meant by utility as a good, while symbolic value (i.e., image) is the extent to which a purchase enhances the worth of the person in his own eyes (self-esteem) and in the eyes of others (status). This report is concerned primarily with the symbolic value, or market image, of various brands of three kinds of products.

The marketable value of product symbolic utility has long been recognized in general theoretical formulations such as Veblen's (1953) "conspicuous consumption" and in commercial advertising practices such as "product endorsement." At the same time there have been few attempts to further the quantitative analysis of "market image" as a theoretical construct. This report, stimulated by a larger study of the relation between product purchase and personality factors (Pohlman, 1969), describes a simple opera-

tion for quantifying two aspects of any identifiable image dimension of a product category or brand: (a) the level of the image on the selected dimension and (b) the clarity, or sharpness, of the image.

The description is presented in the context of a single, specific image dimension, *class* (the slang expression used in the study for prestige), as it applies to approximately 30 brands in each of three product categories (automobiles, beers, magazines). Specifically, the hypotheses are evaluated that ratings of class vary as a function of (a) consumer group involved in the ratings, (b) product type, (c) brand/product familiarity.

## METHOD

### Subjects

Three samples were drawn from three different consumer subpopulations. Samples of college men ( $n = 42$ ) and college women ( $n = 25$ ) were drawn from Introductory Psychology classes at Gettysburg College. An adult men's sample ( $n = 24$ ) was drawn from a church group in Hawthorne, New Jersey.

### Independent Variables

The three sample groups constituted three levels of the consumer group classification variable (Hypothesis 1). The product variable (Hypothesis 2) was defined in terms of three product categories, each of which contained approximately 30 brands: automobiles (27), beers (30), and magazines (30). To evaluate Hypothesis 3, familiarity with various brands was defined by the subjects' ratings on a 6-point graphic scale.

### Dependent Variable

The dependent variable was based on the subjects' median ratings of the various brands with reference to their judged "classiness" and, in several cases, familiarity. Two characteristics of the dis-

<sup>1</sup> Requests for reprints should be sent to Samuel Mudd, Department of Psychology, Gettysburg College, Gettysburg, Pennsylvania 17325.

tributions of these median ratings were considered in evaluating the hypotheses under evaluation: the average median across all brands of a given product type and the variability of the brand medians across product type.

### Procedure

To facilitate data collection, a 9-point graphic scale that could be administered to groups was used. The instructions and scale format for the class ratings and familiarity ratings appears below:

#### INSTRUCTIONS FOR CLASS RATINGS

For each brand of automobile (or beer or magazine) check the category below according to what *you feel is its "class" or status position* in our society. Try to use the whole range of categories in making your judgments so that when you have rated all 30 brands, each of the nine categories has been checked for several brands.

#### Class Scale Descriptors

Extreme low	Very low	Low	Slightly low	Average	Slightly above	High	Very high	Extreme high
1	2	3	4	5	6	7	8	9

#### Instructions for Familiarity Ratings (6-point scale)

For each brand of magazine check the category below according to *how familiar you are* with that particular brand.

#### Familiarity Scale Descriptors

Never heard of	Vaguely familiar	Fairly familiar	Familiar	Familiar	Familiar
Never buy	Never buy	Never buy	Never buy	Occasionally buy	Buy regularly
1	2	3	4	5	6

The scales were administered in the fall of 1970. For each brand and product the median and  $Q$  value of the class-rating distribution were calculated for each of the three samples. Due to the subjects' time limitations in data collection, familiarity ratings could be obtained only for automobiles in the New Jersey sample and only for magazines in the two college samples. No familiarity data were obtained for beers.

### RESULTS

Median class ratings and the variability ( $Q$ ) of those ratings were calculated for each brand by product category and market segment along with comparable values for the familiarity ratings for automobiles and magazines.<sup>2</sup> All hypotheses were evaluated with reference to these data.

Two analyses of variance were conducted, one involving the medians and the other involving the  $Q$  values for the brand-rating distributions. The analysis of variance of median ratings by product type and consumer

group showed that the median class ratings of brands varied significantly ( $F = 19.68, df = 2/170, p < .01$ ) as a function of the consumer group doing the rating and of the interaction of the consumer group variable with product type ( $F = 15.58, df = 4/170, p < .01$ ). A Newman-Keuls analysis (Winer, 1962) of this interaction is summarized in Table 1. Inspection of Table 1 shows that the means of the median ratings across all magazines for the three consumer groups did not differ significantly, but that there were differences among the three consumer groups in the other two product categories. College men and women attributed greater class to cars in general than did New Jersey men ( $p < .01$ ), but did not differ significantly from each other. College men rated beer as having significantly higher class ( $p < .01$ ) than did college women and New Jersey men whose mean ratings did not differ significantly.

A similar analysis of variance of the  $Q$  values for brand class ratings as a function of

<sup>2</sup> Copies of these values for all brands are available upon request from the second author.

TABLE I

SUMMARY OF NEWMAN-KEULS TESTS FOR DIFFERENCE  
BETWEEN AVERAGE CONSUMER GROUP MEDIAN  
AT LEVELS OF PRODUCT TYPE

Product	Consumer group		
	N.J. men	College men	College women
Automobile	5.58	5.91 <sub>a</sub>	5.92 <sub>a</sub>
Beer	4.72 <sub>b</sub>	5.75 <sub>b</sub>	4.71
Magazine	5.67 <sub>c</sub>	5.62 <sub>c</sub>	5.76 <sub>c</sub>

Note. For each product, cells which share a common subscript are not significantly different from each other at the .01 level.

product type and consumer group did not yield significant differences. Inspection of the data revealed that the  $Q$  values for brands within product type varied widely; apparently this within-product variability masked the tendency for the mean  $Q$  value (across brands within category) to vary from one product class to another. When the variability of these means was analyzed alone (based on overall mean  $Q$  value by consumer group and product type), eliminating the effects of brand variability, the product-type effect approached significance and, when the mean product-type  $Q$  values from the original Pohlman study were incorporated into the analysis, the product effect was significant ( $F = 17.86$ ,  $df = 2/6$ ,  $p < .01$ ). Follow-up Newman-Keuls tests showed that the mean  $Q$  value for automobiles (.75) was significantly smaller ( $p < .01$ ) than that for beers (1.04) and magazines (1.06), these latter two being statistically homogeneous.

The evaluation of the relationship of product/brand familiarity to the class variable was based on the pattern of correlations of the medians and  $Q$  values of the rating distributions. These values are shown in Table 2. Correlations between  $Q$  values for class and familiarity were not significant indicating that the dispersion of familiarity ratings was unrelated to the dispersion of class ratings on the same product/brand items. The  $Q$  values from the familiarity ratings were not correlated significantly with the medians of the same items on the class ratings, showing that class image as indicated by median rating was independent of the variability in familiarity

of the brand in the consumer group sampled. Further, the data showed that the median familiarity of a brand in a consumer group was independent of the variability of class ratings ( $Q$ ) of the brand. This finding leads to the conclusion that brand image clarity is unrelated to overall level of familiarity, although the correlation of these two values for automobiles in the New Jersey men's sample ( $r = -.37$ ,  $p < .06$ ) suggests that as familiarity with a brand increases in a sample, the variability of ratings of class decreases.

In both college samples, the correlation was significant between the medians of the familiarity ratings and the class medians for ratings on magazines ( $r = .56$ ,  $p < .01$ ). A

TABLE 2  
INTERCORRELATIONS OF  $Q$ s AND MEDIAN FOR  
CLASS AND FAMILIARITY RATINGS

Group	Automobiles	Magazines
Variability of familiarity ratings vs. variability of class ratings		
College men	—	-.29
N.J. men	.20	—
College women	—	-.22
Variability of familiarity ratings vs. median class ratings		
College men	—	-.16
N.J. men	-.01	—
College women	—	-.08
Median familiarity ratings vs. variability of class ratings		
College men	—	.17
N.J. men	-.37*	—
College women	—	-.14
Median familiarity ratings vs. median class ratings		
College men	—	.56**
N.J. men	.04	—
College women	—	.56**

\*  $p < .07$ ,  $df = 25$ .

\*\*  $p < .01$ ,  $df = 28$ .

TABLE 3  
CORRELATION OF MEDIAN AND Qs ON CLASS  
AND FAMILIARITY SCALES

Group	Autos	Beers	Magazines
Medians vs. Qs on Class Scale			
College men	-.71**	-.13	.05
N.J. men	-.51**	-.33	.53**
College women	-.47*	.12	-.29
Medians vs. Qs on Familiarity Scale			
College men	—	—	-.44*
N.J. men	-.17	—	—
College women	—	—	-.29

\*  $p < .05$ .

\*\*  $p < .01$ .

comparable correlation between these two variables, however, was not found to be significant for the New Jersey men's sample that rated automobiles. This discrepancy is not easily rationalized since the product and the consumer group variables were confounded, both differing from college students and magazines to New Jersey men and automobiles.

Of interest here also is the pattern of correlation coefficients showing the relation of the median ratings to the variability of the ratings on each scale. Table 3 shows that except for the significant positive correlation between the Qs and medians for the ratings of magazines by New Jersey men, the median class ratings tended to be negatively correlated with the variability of those ratings. That pattern was most striking in the ratings of automobiles where there was a significant correlation for all three consumer groups. The negative correlation indicates that as the variability in ratings of the class of the automobiles decreased (greater agreement in ratings of class) the median class value assigned by the raters increased. In other words, the higher the class of the automobile, the more agreement there was on its class (i.e., prestige) value.

Table 3 also lists the intercorrelations obtained between the two statistics for the familiarity-rating distributions. Once again the pattern is toward a negative relationship,

although only the correlation for college men rating magazines was significant. High familiarity tended to be associated with low variability in the familiarity ratings. In other words, as the overall level of familiarity with the product increased, the amount of individual differences in familiarity tended to decrease.

## DISCUSSION

All experimental hypotheses were supported by the rating data. Hypothesis 1, that ratings of the class of particular types of products vary as a function of market segment, was supported by the finding that the three consumer groups rated differently the class of items in two of three product types: Magazines were rated the same overall by all consumer groups, but automobiles and beers were rated differently (New Jersey men attributed less class to cars than did both college groups, and college men attributed more class to beer than the other two groups). The variability of the ratings of the three product types was found to be independent of consumer group. Here the explanation seems to lie in the fact that there was great variability in the ratings of class from brand to brand within each product type. Accordingly, image clarity, as defined by the variability of class ratings, can be said to vary considerably from brand to brand for these three product types. The present study was not designed to get at the determinants of this interbrand variability (except for the exploratory work with the familiarity variable which is discussed below), but it seems clear that the specification of such determining variables would be useful in further work.

Hypothesis 2, concerned with the effects of product type on the variability of "classiness" ratings, was not supported by the initial analyses of variance of the medians and Q values for the ratings. Again, it was obvious that within-product variability in ratings due to brand differences operated to mask a product-type effect. When the average Q value for each product and market segment were analyzed (eliminating brand within-product-type variability) along with the original Pohlman Q values the product effect was

demonstrated. The mean variability of the  $Q$  values for automobiles was 25% less than the mean variability of  $Q_s$  for both beers and magazines. This fact was interpreted to mean that the class image of a typical automobile was less ambiguous than the class image of a typical magazine or beer. This is not particularly surprising since the automobile has long been marketed as a status commodity. The notable thing here is the fact that a fairly simple psychometric procedure provides the investigator with a means to quantify the sharpness, or clarity, of the status image.

Hypothesis 3 dealt with the relationship between class rating and brand/product familiarity. It was expected that differential familiarity with a brand or product type would be associated with differences in class attributed to the various items rated. The data here were incomplete in that magazines and automobiles only were involved in the familiarity ratings. Considering all possible pairs of correlations between medians and  $Q_s$  versus familiarity and class (a total of 12 correlations), only three approached significance. It is premature to interpret these correlations at this time since the pattern is not complete, but the partial pattern is suggestive of the complex way that familiarity can operate to influence class image ratings. Further studies in this area should be designed to sample familiarity effects across all categories of market segment and product type to control for confounding of these two variables.

As a final note, one of the more interesting possibilities in the application of this technique to the analysis of market images is the monitoring of change in image clarity as a function of marketing efforts. There is every reason to expect that both the median and  $Q$  values for any defined dimension of an image will change over time as various marketing forces exert their influence. A special case of this sort of study would involve the assessment of the image of various political candidates throughout the course of their election campaigns.

#### REFERENCES

- COLEMAN, R. P. The significance of social stratification in selling. In H. Barksdale (Ed.), *Marketing in progress*. N.Y.: Holt, Rinehart & Winston, 1964.
- GUILFORD, J. P. *Fundamental statistics in psychology and education*. (4th ed.) N.Y.: McGraw-Hill, 1965.
- POHLMAN, A. Is product purchase related to self-concept? Paper presented at the Fourth Annual Psi Chi Colloquium, Gettysburg College, May 1969.
- SOMMERS, M. S. Product symbolism and the perception of social strata. In J. U. McNeal (Ed.), *Dimensions of consumer behavior*. (2nd Ed.) N.Y.: Appleton-Century-Crofts, 1969.
- TUCKER, W. T., & PAINTER, J. J. Personality and product use. In J. U. McNeal (Ed.), *Dimensions of consumer behavior*. (2nd Ed.) N.Y.: Appleton-Century-Crofts, 1969.
- VEBLEN, T. *The theory of the leisure class*. N.Y.: New American Library (Mentor Book), 1953.
- WINER, B. J. *Statistical principles in experimental design*. N.Y.: McGraw-Hill, 1962.

(Received September 3, 1971)

## EFFECTS OF HUMAN MODELS ON PERCEIVED PRODUCT QUALITY

RABINDRA N. KANUNGO<sup>1</sup> AND SAM PANG

*McGill University, Montreal, Canada*

An experiment was conducted to study the effects of human models in advertisements on the individual's perception of and attitude toward the product. Subjects were tested under three experimental and one control conditions for each of four different products. In the three experimental conditions a male, a female and a male-female pair were used as models. In the control condition the product was presented without any model. The results revealed that the "fittingness" of the models for the product is an important variable in product advertisements. The implications of congruity theories for product advertisements are discussed.

The use of human models in printed advertisements is perhaps as old as advertising itself, but it is much more prevalent now as compared to the early years of this century. Klapp (1941) figures an increase in the use of human models in advertisement from 22% in 1900 to 74% in 1940. The same trend appears to be still in vogue. The main reason for the increasing use of human models may lie in the fact that they provide a more meaningful social context for the product in advertisements, thus arousing more emotional and attitudinal reactions from the consumers toward the product, suggesting that the consumers pay greater attention to the advertisement. While the above explanation appears intuitively plausible, it leaves many questions unanswered. For example, what differences in the attitudes of the consumers are expected when one compares the effects of male and female models in the product advertisements? Is it true as David Ogilvy (1963) asserts that "when you use a photograph of a woman, men ignore your advertisements, and when you use photographs of a man, you exclude women from your audience [p. 148]"? Ogilvy's assertion seems to have been supported by Rudolph (1947), but does this imply that the assertion be true for all kinds of product without qualification? If not, what are the limiting factors? What kind of product goes best with what kind of model to create favor-

able consumer attitudes and why? These are the questions that need to be answered through systematic experimental exploration. Often many advertisers use a male or female model to pose beside the products with the belief that these human models would at least make the product more attractive to the potential customers. Such beliefs are largely based on the advertiser's intuitive feelings rather than systematic and controlled empirical evidence. In fact, very few experimental studies have addressed themselves to this issue (e.g., Smith & Engel, 1968) and many more need to be conducted. The present investigation was dictated by the need for experimental research in the area and was designed to study the effects of variation in the use of human models in advertisements on perception of and attitude toward the products.

### METHOD

#### *Selection of Products*

Four products were chosen for use in the study: a medium priced two-door hardtop car, a medium priced fair-sized sofa, an expensive stereo set with two speakers, and an ordinary 16-inch black and white television set. These were chosen because they are consumer durables, frequently advertised in magazines.

#### *Selection of Attitude Measuring Device*

In order to obtain some meaningful measure of the effects of human models on the attitudes of potential consumers towards the products, it was necessary to determine the most salient features of the products in terms of which consumer attitudes are best expressed. For this purpose, 10 adult male

<sup>1</sup> Requests for reprints should be sent to R. N. Kanungo, Faculty of Management, McGill University, 1001 Sherbrook Street West, Montreal 110, Canada.

and 10 adult female students of McGill University were asked the following question: "What product qualities or characteristics would you look for when buying each of the following products?" Then they were presented with descriptions of the four products listed earlier, one at a time, and their responses were recorded.

From the analysis of their responses, 11 most frequently cited qualities of the car, and 8 most frequently cited qualities of each of the other three products were selected for use in the study. For each of the products, a rating sheet was prepared on which a 7-point unipolar least to most type scale appeared under each relevant product quality. For each scale, the seven points were represented by numbers 1 to 7 with the words "least" and "most" printed on the top of 1 and 7, respectively. The rating sheets were used to measure the attitude toward the product on relevant quality dimensions.

### *Preparation of Advertisements*

Pictures of four young models, two male and two female, and pictures of the four products were cut out from popular magazines and department store catalogues. Care was taken to match the attire and pose of one male model (M) with the other (M'). Likewise, the female models (F and F') were also matched in their attire and pose. The pictures of each of the products were mounted on 11 × 15 inch cardboards. Separate photographs of model M posed an inch to the right of each of the four mounted products were taken. In the same fashion, separate photographs of each product with models M', F, F', MF, and M'F' were also taken. Besides, each product was photographed without any model to serve as control (C) treatment. Each of the photographic prints of 8 × 11 inch size was mounted on a plastic cardboard for presentation to the subjects.

### *Subjects*

Thirty-two male and 32 female students served as subjects in the experiment. Both the male and female subjects were selected at random. The ages of male subjects ranged from 18–30 years with a median age of 23 years, and those of the female subjects ranged from 18–25 with a median age of 20 years.

### *Experimental Design and Procedure*

The design involved three experimental and one control treatment of each product. The three experimental treatments consisted of each product being presented with a male model (M or M'), a female (F or F'), and a male-female pair (MF or M'F'). The control treatment consisted of the product being presented alone without any model.

The male subjects were randomly divided into four groups of eight subjects in each. Each group was exposed to an advertisement of each of the four products in such a manner that the same model was not exposed to a subject more than once. For example, the subjects in Group I were presented with the advertisements of the four products in the fol-

	MODELS				C
	M or M'	F or F'	MF or M'F'		
Group I	Car	Sofa	Stereo	T.V.	
Group II	Sofa	Car	T.V.	Stereo	
Group III	Stereo	T.V.	Car	Sofa	
Group IV	T.V.	Stereo	Sofa	Car	

FIG. 1. Experimental design indicating exposure of product advertisements to four groups of male and female Subjects.

lowing manner: the car with a male model (half of the subjects getting M and the other half getting M'), the sofa with a female model (half of the subjects getting F and the other half getting F'), the stereo with a male-female pair (half of the subjects getting M'F' and the other half getting MF), and the television without any model (C). Similarly, subjects in the other three groups (Groups II, III, and IV) were exposed to the advertisements of the four products in a manner as shown in Figure I. Exactly the same design was followed for the 32 female subjects. It may be noticed that for each product, the design provides for four treatments with independent male and female groups of subjects.

The subjects were tested individually. Each subject was presented with a set of four product advertisements, one at a time, depending on the group to which he or she was assigned. With the exposure of each advertisement, the subject was instructed to evaluate the product appearing in the advertisement by using the appropriate rating sheet. The rating sheet contained the 7-point scales representing relevant qualities, and the subject indicated his or her rating by circling the appropriate points on the scales. The order of presentation of the advertisements was randomized for each subject. It took approximately 12–15 minutes for each subject to rate all four products.

### **RESULTS**

The quality ratings of each product by male and female subjects were analyzed separately for the purpose of comparison. In addition, for each product, mean ratings of each experimental treatment were compared with those of the control treatment.

#### *Analysis of Perceived Qualities of Car*

Comparisons of mean quality ratings of the car with and without models as reflected in mean difference scores<sup>2</sup> are presented separately for male and female subjects in Table 1. It may be noted that the car was rated on

<sup>2</sup> Complete data can be obtained from the authors on request.

TABLE 1  
COMPARISONS OF MEAN RATINGS OF THE CAR WITH AND WITHOUT MODEL  
AS REFLECTED IN MEAN DIFFERENCE SCORES

Product qualities	Treatments compared with control					
	Male model		Female model		Male-female pair	
	Male ratings	Female ratings	Male ratings	Female ratings	Male ratings	Female ratings
Gas economy	00	+0.25	-0.25	+0.87**	-0.37*	+0.12
Accommodative	+0.62**	+0.25	+0.75**	+0.62**	+0.12	-0.13
Comfortable	-0.12	-0.37	-0.50	-0.62*	00	+0.63**
Safety	-0.50*	+0.37*	-0.62*	+1.00**	-0.25	+0.50*
Usefulness	+1.73**	-0.88**	+1.00**	00	+1.75**	-0.50*
Lively	+1.00**	+1.13**	-0.25	-0.12	+0.62**	+0.13
Appealing	+1.13**	+0.63*	-0.75**	-0.75*	-0.12	-1.12**
Fashionable	+0.75**	+0.62**	-0.25	-1.00**	-0.50	+0.12
Horsepower	+0.37*	+0.12	-0.75**	-0.88**	+0.12	-0.13
Strength	-0.63*	+0.25	-0.75**	+0.25	-0.75*	-0.37
Easy to handle	-0.25	-0.25	-1.13**	-0.50*	-0.13	-0.50*

\* $p < .05$  (two-tailed test).

\*\* $p < .01$  (two-tailed test).

11 different product quality dimensions. For each dimension the mean rating of the car without any model (control) served as the base line against which the mean rating of the car with the model (experimental) was compared. In Table 1, a positive mean difference score indicates a higher rating for the experimental compared to the control treatment, and a negative score indicates the reverse.

Inspection of Table 1 reveals that, compared to the control treatment, the car with a male model causes among male subjects significant favorable impressions (positive mean differences) on 6 out of 11 product qualities and significant unfavorable impressions (negative mean differences) on only 2 out of 11 product qualities. On the remaining three product qualities there was no significant change. Similar comparisons for female subjects reveals that introduction of a male model caused significant favorable impressions on four and unfavorable impressions on only one product quality. On the remaining six product qualities there were no significant changes. The findings suggest that for both male and female subjects, the male model has more positive than negative effects. It causes greater favorable than unfavorable product images.

With a female model, the car received significantly favorable evaluations on only two, but unfavorable evaluations on five product qualities from male subjects. Likewise, it received significantly favorable evaluations on three, and significantly unfavorable evaluations on five product qualities from female subjects. This suggests that the use of a female model causes a more unfavorable than favorable image of the car for both male and female subjects.

From male subjects, the car with a male-female pair received significantly favorable evaluations on two and unfavorable evaluations on another two product qualities. Similarly, female subjects gave significantly favorable ratings only on two and unfavorable ratings on three product qualities. This indicates that use of a male-female pair with the car has no distinct advantage over the control for both male and female subjects.

It seems that from the advertiser's point of view, the use of a male model with the car would perhaps best accomplish the aim of creating a favorable product image. On the other hand, the use of a female model should be avoided because of its greater unfavorable than favorable effects on the product image.

TABLE 2  
COMPARISONS OF MEAN RATINGS OF THE SOFA WITH AND WITHOUT MODEL  
AS REFLECTED IN MEAN DIFFERENCE SCORES

Product qualities	Treatments compared with control					
	Male model		Female model		Male-female pair	
	Male ratings	Female ratings	Male ratings	Female ratings	Male ratings	Female ratings
Comfortable	-0.87**	-0.37*	+0.50**	+0.38	-0.50*	+0.13
Usefulness	-1.38**	-0.25	+0.12	-0.62**	-0.63**	+0.25
Fashionable	-0.25	-0.63**	+1.00**	+0.50*	+0.75*	+0.50*
Strength	-0.63*	-0.38	+0.50*	+0.12	-0.50*	+0.12
Softness	-0.25	-1.00**	+1.00**	-1.00**	+0.33	-0.12
Decorative	-0.13	00	+1.25**	+0.12	-0.25	00
Durability	-0.37*	+0.62**	+0.13	+0.37**	-0.50**	00
Large	-0.25	-1.13**	+0.75**	-0.13	+0.13	-0.88**

\*  $p < .05$  (two-tailed test).

\*\*  $p < .01$  (two-tailed test).

#### *Analysis of Perceived Qualities of the Sofa*

Table 2 presents difference scores derived from mean quality ratings of the sofa with and without a model. Compared to the control, the sofa with a male model received significantly unfavorable evaluations on four out of eight product qualities from both male and female subjects. However, it received significantly favorable evaluation only on one product quality and that too only from the female subjects.

With the use of a female model, the sofa received significantly favorable evaluations on six out of eight product qualities from male subjects. Female subjects however, gave significantly favorable evaluations on two and unfavorable evaluations on another two product qualities.

Compared to the control, the sofa with a male-female pair received significantly unfavorable evaluations on four and favorable evaluations only on one product quality from male subjects. From female subjects, it received significantly favorable and unfavorable evaluations on one product quality each.

The above findings suggest that the advertiser would be better off using a female model for the sofa because it creates a more favorable than unfavorable product image, although the effect is limited only to male subjects. The use of a male model for the sofa

should be avoided because it causes greater harm than good to the image of the sofa in the minds of both male and female subjects. A male-female pair also does not seem to provide a more favorable image of the sofa in either male or female subjects.

#### *Analysis of Perceived Quality of the Stereo*

Table 3 presents difference scores reflecting comparative evaluation of the stereo with and without a model. Compared to the control, the stereo with a male model did not receive any significant unfavorable evaluations by either male or female subjects. On the other hand, it was significantly favorably evaluated on two and five product qualities, respectively, by male and female subjects.

It will be noticed that, compared to the control, the stereo with a female model received significantly favorable evaluations on five and unfavorable evaluations on only one out of eight product qualities from both male and female subjects.

The stereo with a male-female pair received significantly favorable evaluations on five product qualities from male subjects and on four product qualities from female subjects. Only the latter group judged it significantly unfavorably on only one product quality.

The above results indicate that the favorable image of the stereo is enhanced with the

TABLE 3

COMPARISONS OF MEAN RATINGS OF THE STEREO WITH AND WITHOUT MODEL  
AS REFLECTED IN MEAN DIFFERENCE SCORES

Product qualities	Treatments compared with control					
	Male model		Female model		Male-female pair	
	Male ratings	Female ratings	Male ratings	Female ratings	Male ratings	Female ratings
Sound effect	-0.12	+0.50**	+1.00**	+0.38*	+0.50*	+0.88**
Strength	-0.50	+0.38	-0.62*	+0.13	-0.25	-0.12
Attractive	-0.12	+0.88**	+0.75*	+0.76**	+0.75*	+0.63**
Fashionable	+1.00**	+1.25**	+1.00*	+0.50	+0.62**	+0.75
Durability	.00	+0.75**	-0.13	+0.63**	.00	+0.87**
Decorative	+0.38	+0.75**	+1.00**	+0.50*	+0.25	+0.37
Usefulness	-0.12	-0.13	+0.50*	+0.37*	+1.37**	+0.50**
Easy to handle	+0.53**	+0.37	+0.25	-1.38**	+1.38**	-0.63*

\*  $p < .05$  (two-tailed test).

\*\*  $p < .01$  (two-tailed test).

use of any of the three variations of human models. Both male and female subjects responded more favorably to the stereo when it was presented with a model than without a model.

#### *Analysis of Perceived Qualities of the Television*

Table 4 shows the effects of the models on the image of the television. Compared to the control, the television with a male model was

significantly unfavorably evaluated on six and favorably evaluated only on one out of eight product qualities by male subjects. However, female subjects evaluated it significantly unfavorably on two and favorably on another three product qualities.

Male subjects gave significantly unfavorable evaluations of the television on five and favorable evaluations on one of the eight product qualities. Female subjects however, evaluated it significantly favorably on two

TABLE 4

COMPARISONS OF MEAN RATINGS OF THE TELEVISION WITH AND WITHOUT MODEL  
AS REFLECTED IN MEAN DIFFERENCE SCORES

Product qualities	Treatments compared with control					
	Male model		Female model		Male-female pair	
	Male ratings	Female ratings	Male ratings	Female ratings	Male ratings	Female ratings
Sound effect	-0.50*	.00	-0.50*	-0.13	+0.25	+0.12
Strength	-1.25**	.00	-1.50**	-0.50**	-1.75**	-0.37*
Attractive	+0.25	+0.88*	+0.25	+0.75**	+0.25	+0.50**
Durability	-0.88**	-0.63**	-0.75**	.00	-0.75**	+0.25*
Decorative	-0.63*	+0.37*	+0.37	+0.25	.00	+0.25
Usefulness	-0.87**	1.62**	-1.75**	+0.13	-1.38**	-0.12
Easy to handle	-1.38**	-0.50	-0.63*	+1.00**	-0.88**	+0.63*
Reception	+0.50*	+0.63*	+0.87**	+0.38	+0.12	+0.63**

\*  $p < .05$  (two-tailed test).

\*\*  $p < .01$  (two-tailed test).

and unfavorably on one out of eight product qualities.

The television with a male-female pair was evaluated by male subjects significantly unfavorably on four product qualities. On the other hand, female subjects evaluated it more favorably on the same number of product qualities, and only on one product quality, did they rate it unfavorably to a significant extent. It appears that for male subjects, presence of any model creates a more unfavorable impression of the television. However, in the case of female subjects, a male-female pair seems to create a more favorable impression of the product.

### DISCUSSION

Overall, the findings of the study reveal that the effects of human models on consumer attitude is not as simple as Ogilvy's (1963) assertions. There appears to be an interaction between the nature of the product and human models. The three variations of human models in this study seems to have differential effects for different products. Use of a model with one product may cause a favorable attitude toward the product, whereas use of the same model with another product may cause an unfavorable attitude. The study revealed that an overall favorable attitude toward the product was created in both male and female subjects when a male model was used for the car and when a female model was used for the sofa. On the other hand, a female model for the car and a male model for the sofa created more unfavorable attitude toward the product. In the case of the stereo, generally favorable attitude resulted from the use of any of the three model variations: male, female, and male-female pair. However, presence of the same models caused unfavorable attitude toward the television among the male subjects. The female subjects responded quite favorably to the television, only when a male-female pair was used.

Why should a particular product with a male model be viewed favorably whereas another product with the same model be viewed unfavorably? What determines the product-model interaction? The answers may lie in the "fittingness" of the model for the product. Each product, when perceived, evokes some

general or stereotype image depending on its features and the associations it brings to our mind. Thus, some products are perceived as either being predominantly masculine or predominantly feminine. Some products may even be perceived as neither. It is proposed that the fittingness of a male model is greater for a product with a masculine image than for a product with a feminine image. Likewise, the fittingness of a female model is greater for a product with feminine image than for a product with masculine image. Whenever an advertisement provides such fittingness or product-model match, the person exposed to the advertisement experiences perceptual and attitudinal congruity. Such congruity perhaps results in his increased favorable attitude toward the product because congruous experience is psychologically comfortable for the individual (Zajonc, 1960). On the other hand, when the person is exposed to a non-fitting or product-model mismatch type of advertisement, he experiences perceptual and attitudinal incongruity. Experiencing an incongruous situation is psychologically uncomfortable and hence, the experience expresses itself in an increased unfavorable attitude toward the product.

TABLE 5  
PERCENTAGES OF SUBJECTS EVALUATING THE PRODUCTS IN TERMS OF FOUR ANSWER CATEGORIES

Product	Sample	Answer categories			
		Masculine	Feminine	Both	Neither
Car	Male	73	7	13	7
	Female	80	5	10	5
Sofa	Male	—	80	20	—
	Female	—	60	20	20
Stereo	Male	7	7	60	26
	Female	20	20	40	20
Television	Male	7	20	—	73
	Female	5	—	30	65

Note. Decimal points are omitted.

In order to test the fittingness hypothesis in the context of the present study, 20 female (age range 18-29 years) and 15 male (age range 18-30 years) students were asked to categorize each of the four product pictures without any model into the following categories: (a) appears more masculine than feminine, (b) appears more feminine than masculine, (c) appears equally masculine and feminine, and (d) appears neither masculine nor feminine. The percentages of responses for each of the four products are presented in Table 5. In general, the results in Table 5 substantiate the fittingness hypothesis. Most of the subjects judged the car as masculine, the sofa as feminine, the stereo as both masculine and feminine, and the television as neither. Considering the results of the study, this is what one would expect based on the fittingness hypothesis.

Questions may be raised regarding the generalizability of the present findings. Should male models be used and female models be avoided in advertisements of all kinds of cars? Obviously not. In advertisements of some cars, the product may appear as feminine

depending on its color, size, shape, and other features. For these advertisements, a female model is perhaps a better fitting one than a male model. Thus, it is important to ascertain the perceived qualities of the product picture first before deciding what kind of model to use. The primary consideration, however, should be to obtain a best match between the product and the model.

#### REFERENCES

- KLAPP, O. E. Imitation value in advertising. *Journal of Applied Psychology*, 1941, 25, 243-250.
- O'CEILVY, D. *Confession of an advertising man*. New York: Dell, 1963.
- RUDOLPH, H. J. *Attention and interest factors in advertising*. New York: Funk & Wagnalls & Printer's Ink, 1947.
- SMITH, G. H., & ENGEL, R. Influence of a female model on perceived characteristics of an automobile. *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, 3, 681-682.
- ZAJONC, R. B. The concepts of balance, congruity and dissonance. *Public Opinion Quarterly*, 1960, 24, 280-296.

(Received September 27, 1971)

## PERSONALITY AND PRODUCT USE REVISITED: AN EXPLORATION WITH THE PERSONALITY RESEARCH FORM

PARKER M. WORTHING<sup>1</sup>

*University of Massachusetts*

M. VENKATESAN

*University of Iowa*

STEVE SMITH

*Dartmouth College*

Past studies relating personality traits and product usage patterns have utilized personality tests that may have been inappropriate inasmuch as they have been developed for specialized and sometimes diagnostic purposes. This exploratory study used scales from Jackson's (1967) *Personality Research Form* (PRF), which is intended for a wide variety of situations, including consumer behavior. Undergraduate college students ( $n = 232$ ) completed five scales of the PRF and also indicated product usage information. Canonical analysis was used to analyze the results of this study. The findings confirm the complex relations between personality traits and product use.

Interest in the relationships between personality variables and product usage or brand preference remains strong among consumer behavior researchers. However, attempts to identify and substantiate such relationships have not been promising, (see e.g., Evans, 1959; Westfall, 1962; Marcus, 1965), although some significant relationships have been found (e.g., Cohen, 1967; Gottlieb, 1959; Jacobson & Kossoff, 1963; Koponen, 1960; Tucker & Painter, 1961).

Limitations in comparing earlier studies are the variety of instruments utilized, variations in product category definitions, usage rate classifications and brand selections. Even a cursory comparison of selected product categories (e.g., automobile or cigarettes) that have been found to be associated with personality traits (e.g., sociability, emotional stability, ascendancy, and the like) reveals inconsistencies. Even studies (such as Koponen's, 1960) utilizing product usage data that came from the same panel of consumers and reflecting the same personality scores have provided disappointing results (Advertising Research Foundation, 1964; Massy, Frank, & Lødahl, 1968). The result has been an emerging body of critical appraisal of previous research efforts and suggestions for improved

methodological and analytical approaches for subsequent research efforts.

There are two major shortcomings that have been underscored recently with respect to the studies dealing with personality traits and product purchase (or usage) behavior. First, the use of personality instruments such as the ones indicated earlier on a specific population in a consumer behavior context is inappropriate inasmuch as these instruments were originally developed for specialized uses and diagnostic purposes far removed from situations involving consumer behavior (Brody & Cunningham, 1968; Kassarjian, 1971; Wells, 1966). Too little a priori thought is given to how and why personality should or should not be related to given aspects of consumer behavior (Jacoby, 1971).

Secondly, Sparks and Tucker (1971) point to an analytical weakness in previous studies. The usage of bivariate inferential techniques and regression including multiple correlation implies that personality is comprised of a packet of discrete, independent traits which do not interact or exert interrelated influences on one's product or brand preferences. They have suggested that analytical shortcomings can be alleviated by the use of canonical analysis, which can synthesize individual personality traits into molar personality types. Their study found that canonical analysis provided further insight into the

<sup>1</sup> Requests for reprints should be sent to Parker M. Worthing, University of Massachusetts, Amherst, Massachusetts 01002.

complexity of personality and product usage relationships.

From the foregoing, it is apparent that there is a need for selection of instrument(s) relevant to the context of consumer behavior. A review of previous studies revealed that five traits stand out in terms of their relationships to product usage patterns: affiliation, aggression, dominance, exhibition, and social recognition. The present study utilizes an instrument<sup>2</sup> designed to measure each of these five traits. Specifically, the objective of this study was to explore relations between these general personality traits and product usage.

#### METHOD

The aforementioned five scales, each consisting of 20 items, were administered to 232 college students from the University of Massachusetts and the University of New Hampshire. There were 166 males and 66 females. One hundred forty four subjects were from introductory marketing classes, and 88 were from introductory psychology classes.

The subjects came to the laboratory as part of the course requirement to participate in a research project. They were given the 100-item personality questionnaire. Following completion of this questionnaire, they were given a product usage questionnaire to fill out. The product usage questionnaire contained 18 product categories, with which most college students were expected to be familiar. In fact, the categories were arrived at after examining the product categories that had been used by the previous studies. The subjects were asked to check either "Yes" or "No" for usage of each product in the questionnaire.

The administration of the five scales and the product questionnaire was completed in a single sitting simultaneously at both universities, so that there was no possibility of contaminating communication among subjects.

#### Replication

The same five scales were administered to a second sample of 133 male subjects, who were all volunteers from basic courses in business administration at the University of Massachusetts. For this replication, the same procedures were followed.

<sup>2</sup> The PRF is a rational inventory (Edwards, 1970), consisting of 14 traits and one validity scale in the short form (Forms A and B), and the extended form (Forms AA and BB) contains 20 personality variables and two validity scales. Theoretical considerations and factor analytic results have led to the organization of the PRF into six units. Two of these units—measures of degree of ascendency and measures of degree and quality of interpersonal orientation—contain eight scales, including the five traits of interest.

#### RESULTS AND DISCUSSION

One major concern, which has received too little attention in previous studies, is with the reliability and validity of the instrument when it is used with samples that are different from the ones on which the instrument is validated. For example, norms for many of the personality instruments are based on samples of college students (mostly students of psychology). A measure of reproducibility of the scales is needed to satisfy that the instrument is reliable when it is used with a diverse group such as consumers or students in business administration and the like.

A simple comparison of the means and standard deviation of our results to the test norms provides a measure of reproducibility to our sample. Jackson (1967) indicated that based on large samples of college students, Personality Research Form (PRF) was standardized and the normative scores were reported to have a mean of 50 and a standard deviation of 10 for both the males and females. Means and standard deviations for the standardized scores of the respondents in this study for the five scales are indicated in Table 1. It is clear from Table 1 that our sample, consisting of a major share of nonpsychology students, obtained scores on these five scales similar to the scores obtained by the normative samples used by Jackson (1967).

A second concern is with the independence of the scales. In most of the previous studies dealing with traits-product use relationships, no attempt has been made to check the independence of the scales used and where such attempts have been made, independence of scales has not been observed (Sparks & Tucker, 1971; Tucker & Painter, 1961). Jackson (1967) indicated that correlations between PRF scales were generally low or moderate, indicating that each scale possessed substantial unique variance. Discrimination among respondents on the basis of differences among traits can be sharpened, if the studies that use the scales can reproduce such reduced correlations between scales. As a check on the discriminant validity of the scales used in this study, intercorrelations between the five scales were obtained and compared with the low intercorrelations obtained by Jack-

TABLE 1  
MEANS AND STANDARD DEVIATIONS FOR  
STANDARDIZED SCORES

Factor	Males (n = 166)		Females (n = 66)	
	M	SD	M	SD
Affiliation	53.60	10.12	51.53	10.24
Aggression	49.11	10.30	50.36	9.52
Dominance	50.57	11.74	49.41	11.54
Exhibition	48.72	9.60	47.92	10.41
Social recognition	48.93	10.40	50.01	11.72

son (1967). The comparison can be seen in Table 2, where the intercorrelations obtained by Jackson are indicated above the major diagonal and those that were obtained in this study are shown below the major diagonal; comparison of these intercorrelations indicate similar results of independence for these five scales.

To discern differences in personality traits

TABLE 2  
INTERCORRELATIONS AMONG PERSONALITY TRAITS

Traits	1	2	3	4	5
Affiliation (1)		-.09*	.27	.34	.27
Aggression (2)	-.01		.18	.31	.29
Dominance (3)	.25		.34	.54	.17
Exhibition (4)	.40		.40	.52	.31
Social recognition (5)	.37		.30	.18	.34

\* Correlations above the major diagonal are those obtained by Jackson (1967), while those below the major diagonal are those obtained in the present study.

between users and nonusers of these 18 product categories, a one-way analysis of variance was used. Separate analyses were made for each of the five scales with each of the 18 product categories. Analyses for females included only 11 product categories as there were too few nonusers for remaining product categories. Thus out of a total of 55 possible comparisons, only 5.4% of the *F* values were found to be significant at the 5% level. Since

TABLE 3  
ANALYSIS OF VARIANCE AND CORRELATIONS OF PRODUCT USE AND  
PERSONALITY TRAITS (MALES: n = 166)\*

Product	User	Nonuser	Affiliation	Aggression	Dominance	Exhibition	Social Recognition
Cigarettes	70	96	6.15**	19.60**	4.10**	14.81**	6.80**
Razor blades	127	39	0.89	.33*	.16	.29*	.08
Electric shavers	93	73	.07	-.02	-.02	.10	0.03
Radios	151	15	0.68	0.48	0.21	0.15	-.01
Beer	142	24	.06	.14	.14	.13	1.31
Soft drinks	161	5	6.23*	0.63**	7.81**	5.80*	4.15*
Headache remedies	133	33	-.01	.22*	.08	.17	.06
Mouthwash	123	43	3.66	8.36**	4.22*	3.35	0.02
Deodorant	163	3	.15	.22*	.16	.14	.01
Mens' cologne	134	32	8.70**	0.96	3.03	2.53	4.13*
Mens' after shave	137	29	1.17	6.51*	2.34	4.93*	3.01
			.08	.19*	.12	.17	.13
			.22*	0.52	3.78	2.31	0.00**
			.22*	.06	.15	.12	.23*
			.22*	.00	2.24	1.31	2.13
				.00	.12	.09	.11

\* For product categories not indicated here, none of the *F* values or correlations were significant.

b For each product, the first row indicates the *F* values obtained from the one-way analysis of variance.

c For each product, the second row indicates the point-biserial correlations.

\*  $p < .05$ .

\*\*  $p < .01$ .

TABLE 4  
RESULTS OF CANONICAL ANALYSIS

Variables	Canonical coefficients			
	First study	Replication		
Predictor Set (Personality traits)	1	2	1	2
Affiliation	.53	.59	-.03	.16
Aggression	.76	-.51	-.18	-.64
Dominance	.14	.50	.99	-.54
Exhibition	.08	.57	-.31	.18
Social Recognition	-.11	.45	.58	.84
Criterion Set (Product use)				
Cigarettes <sup>a</sup>	.56	-.34	.16	.15
Razor blades	-.19	.07	.19	-.06
Electric shavers	-.11	.27	-.13	.35
Radios	.15	-.10	-.01	-.04
Beer <sup>a</sup>	.34	.16	-.31	.28
Soft drinks	.19	-.32	-.13	.30
Headache remedies <sup>a</sup>	.32	-.18	-.15	-.01
Mouthwash <sup>a</sup>	.95	-.55	.22	-.24
Deodorant	.16	.00	.03	-.47
Mens' cologne	.19	.31	.05	.33
Mens' after-shave <sup>a</sup>	.10	.33	.35	.12
Mens' hair dressing	.00	.10	-.11	.14
Shampoo <sup>a</sup>	-.14	.04	-.47	.27
Handcream	.10	.11	.06	.05
Hair spray	-.12	-.23	.27	.47
Mens' dress shirts	.89	.83	.24	.17
Mens' suits	-.07	.25	.37	.09
Mens' dress shoes	.09	.01	.21	.46
Roots	.34	.18	.38	.17
Canonical R	.58	.43	.61	.41
Chi-square	119.08	54.60	114.9	57.3
Probability	.05	ns	.05	ns
df	90	68	90	68

<sup>a</sup> Products for which Sparks and Tucker (1971) found the canonical index to be related.

5% of the *F* values would be expected to be significant by chance, the results presented here exclude female subjects.

The relationships between the five traits and the user/non-user category, based on the one-way analysis of variance for males, and the point-biserial correlations that were obtained are shown in Table 3. Twenty-one *F* values are significant or about 23% of the obtained values are significant, while only 5% of them can be expected to be significant by chance alone. There are 14 significant correlations where 5 are expected to occur by chance. All 14 relationships were significant at the .05 level or better. It appears that analysis of variance is much more sensitive to potential trait-product use relationships. With the exception of razor blades, the present study corroborates the findings of other studies (e.g., Cohen, 1967; Tucker & Painter, 1961) that there are trait-product use relations, although the particular traits specifically related to a given product category are not necessarily confirmed.

Since Sparks and Tucker (1971) have persuasively argued that individuals' psychological makeup, cognitive or affective, does

not operate as a result of discrete personality characteristics, in order to further understand the relationships among traits and particular product use, canonical analysis was employed to study the kinds of personality structures involved. The results of the canonical analysis are presented in Table 4.

In Table 4, only the first two canonical roots are shown for both the first study and replication; the other three roots derived were far from significant. The first root of the original study with an *R* of .58 is significant at the .05 level of significance. These two roots combined account for 52.5% of the variance. Using an arbitrary value of .3 as a cutoff point for the canonical coefficients of the predictor and criterion variables, some interpretation of the relationships between these traits and product uses can be attempted.

The first root is associated with affiliation and aggression and is related to use of cigarettes, beer, headache remedies, mouthwash and men's dress shirts. The second root is associated with all five traits and is related to the infrequent use of cigarettes, soft drinks, and mouthwash, and use of men's cologne, men's aftershave, and men's dress shirts.

Sparks and Tucker (1971) found sociability, emotional stability, and irresponsibility to be the determinant predictor variables and cigarettes, alcoholic beverages, shampoo and early fashion adoption to be the heaviest loaded factors for the criterion variables. The second root obtained by Sparks and Tucker (1971) showed sociability, cautiousness, and emotional stability as the predictor variables and headache remedies, mouthwash, late fashion adoption, and aftershave lotion as the product loadings.

For the replication, again the first root was found to be significant ( $\chi^2 = 114.9$ ,  $df = 90$ ,  $p < .05$ ), and the general pattern appears similar. That is, 10 product loadings, as compared to 11 for the first study, were above the cut-off point. However, these loadings were not for the same product categories. High loadings are found for only four out of the original seven product categories in the replication. Similarly, the first root, while

significant, is seen to be associated with dominance, exhibition, and social recognition. Some post-hoc explanations are available for these findings. For example, the sample for the replication came from students of one basic required course in business administration. Since the conduct of the first study, usage of some of the product categories among college students might have changed significantly. While the two sets of these results are not entirely similar, both indicate that product usage seems to be related to a complex of personality trait interactions, the nature of which we are barely beginning to understand.

### SUMMARY

One of the major shortcomings of the studies relating personality traits to product usage has been in the use of appropriate instruments. This exploratory study using PRF indicates that this instrument might overcome the limitations associated with other instruments developed for specialized purposes. Our findings corroborate somewhat the findings of other studies and in general it would appear that the scales have provided a measure of convergent validity. The canonical analysis, while illuminating, is difficult to interpret. However, the nature of the interrelations among personality traits and product usage is more clearly revealed by the use of canonical analyses.

While the results of the replication and its analyses did not produce results entirely similar to the first study, a complex interaction between traits and product usage appears indicated. The study suggests the potential usefulness of PRF for prediction of product usage and the utility of canonical and other multivariate analyses to discern complex relationships between personality traits and product usage is evident, as these techniques of data analysis are just beginning to be applied in this area.

### REFERENCES

- ADVERTISING RESEARCH FOUNDATION. *Are there consumer types?* New York: Author, 1964.
- BRODY, R. P., & CUNNINGHAM, S. M. Personality variables and the consumer decision process. *Journal of Marketing Research*, 1968, 5, 50-57.
- COHEN, J. An interpersonal orientation to the study of consumer behavior. *Journal of Marketing Research*, 1967, 4, 270-278.
- CRITES, J. O. Test reviews. *Journal of Counseling Psychology*, 1969, 16, 181-184.
- EDWARDS, A. L. *The measurement of personality traits by scales and inventories*. New York: Holt, Rinehart & Winston, 1970.
- EVANS, F. B. Psychological and objective factors in the prediction of brand choice: Ford versus Chevrolet. *Journal of Business*, 1959, 32, 340-369.
- GOTTLIEB, M. J. Segmentation by personality types. In L. H. Stockman (Ed.), *Advancing marketing efficiency*. Chicago: American Marketing Association, 1959.
- JACKSON, D. N. *Manual for the Personality Research Form*. (Research Bulletin No. 43) London, Ontario: University of Western Ontario, 1967.
- JACOBSON, E., & DOSHOFF, J. Self-percept and consumer attitudes toward small cars. *Journal of Applied Psychology*, 1963, 47, 242-245.
- JACOBY, J. Personality and innovation proneness. *Journal of Marketing Research*, 1971, 8, 244-247.
- KASSARJIAN, H. H. Personality and consumer behavior: A review. 1971. (Mimeo)
- KOPONEN, A. Personality characteristics of purchasers. *Journal of Advertising Research*, 1960, 1, 6-12.
- MARCUS, A. S. Obtaining group measures from personality test scores: Auto brand choice predicted from the Edwards Personal Preference Schedule. *Psychological Reports*, 1965, 17, 523-531.
- MASSY, W. F., FRANK, R. E., & LODAHL, T. M. *Purchasing behavior and personal attributes*. Philadelphia: University of Pennsylvania Press, 1968.
- SPARKS, D. L., & TUCKER, W. T. A multivariate analysis of personality and product use. *Journal of Marketing Research*, 1971, 8, 67-70.
- TUCKER, W. T., & PAINTER, J. J. Personality and product use. *Journal of Applied Psychology*, 1961, 45, 325-329.
- WELLS, W. D. General personality tests and consumer behavior. In J. Newman (Ed.), *On knowing the consumer*. New York: Wiley, 1966.
- WESTFALL, R. Psychological factors in predicting product choice. *Journal of Marketing*, 1962, 36, 34-40.

(Received July 26, 1971)

## DIMENSIONS OF ATTITUDES TOWARD TECHNOLOGY

ROY D. GOLDMAN,<sup>1</sup> BRUCE B. PLATT, AND ROBERT B. KAPLAN

*University of California, Riverside*

An 80-item questionnaire measuring attitudes toward mechanization was administered to undergraduate students in physical science, biological science, social science, and fine arts. Responses were factor analyzed using a varimax rotation. Factor scores were created for six of the resulting factors. These factor scores were then used as dependent variables in a multivariate comparison of the students in different major fields. Most of the between group differences in attitude toward mechanization were reflected by differences in mechanical curiosity.

Contemporary man is witnessing an unprecedented technological boom. Science affects his life in every conceivable way. One salient aspect of contemporary America is its high degree of mechanization—automation and dependency on machines is steadily increasing. Mechanization takes such diverse forms as: nationwide credit card systems, automated assembly lines, and use of computers for scientific research.

There is no shortage of contemplative literature regarding the future of science and technology (see Calder, 1965; Kahn & Wiener, 1967; Prehoda, 1967). Although there are numerous beneficial aspects of new technologies, some feel mechanization is a major source of human grief. By mechanically manipulating the environment man may irreversibly disturb ecological balance. It has also been hypothesized that advancing technology is creating an intolerable acceleration of the pace of life (Toffler, 1970). Mechanization may also mean loss of privacy, accelerated nuclear proliferation, or increases in unemployment (Kahn & Wiener, 1967). These problems are recognized by technology's advocates to be the short-run costs of mankind's long-range profit and, therefore, are acceptable. While scientists and engineers know and describe the workings of machines, it is not their province to decide if machines ought to be used. The costs and benefits are those of society as a whole and, as Krech (1966) maintains, the responsibility of decision rests with society.

In light of the profound effects of technol-

ogy, there is considerable interest in both describing and measuring attitudes toward technology. There may be a number of variables contributing to the individual's dispositions toward technology. These might include: (a) curiosity about machines, (b) alienating effects of a technological society, and (c) perceived quality of machine manufactured goods, as well as others.

If attitudes toward technology are at all predictive of behaviors such as voting or buying, then assessment of such attitudes can be quite useful to society's decision makers. Similarly, scales can serve useful as *dependent variables* in assessing the efficacy of persuasion attempts. In addition, it is likely that attitudes toward technology will be related to attitudes toward the environment in general. It is interesting to note that, while there are several recent studies concerned with attitudes toward the environment (McKechnie, 1970), the professional literature appears devoid of a measure of dispositions toward machinery in particular.

Since various vocations have differential exposure to machinery, we might expect quite different dispositions toward mechanization among members of different professions. For example, research on Strong Vocational Interest Blank (SVIB) has successfully demonstrated a relationship between interests in machines and successful adjustment to a profession involving a high degree of contact with machinery (Campbell, Borgen, Eastes, Johannson, & Peterson, 1968). Similarly we might expect differential dispositions toward technology among college students involved in various disciplines. For example, students

<sup>1</sup> Requests for reprints should be sent to Roy D. Goldman.

involved in the physical sciences have considerably more contact with machines than those in fine arts or humanities.

The purpose of the present research is two-fold. First, it was an attempt to discover different dimensions of attitudes toward some of the different aspects of technology mentioned above as a preliminary to future scale development and, second, to compare college students involved in various disciplines on the basis of these derived measures.

#### METHOD

##### *Questionnaire*

The data consist of responses to an "attitude-toward-mechanization" questionnaire. The 80 items for this questionnaire represent a distillation of a large collection of statements expressing attitudes toward numerous facets of technology. Examples of items in this questionnaire are shown in the results section.

##### *Subjects and Procedure*

The questionnaire was administered to four groups of undergraduate students at the University of California, Riverside. The four groups of students and the number in each group were: fine arts, 54; social sciences, 167; biological science, 177; and physical science, 58. The relative proportion of students in each of these fields is fairly representative of the total number of undergraduates majoring in these fields. It was expected that students in these different major fields would express different attitudes toward mechanization since these fields require different amounts of contact with the processes and products of technology. In particular, it was expected that students in physical science would have the greatest contact with machines, followed by students in biological science, social science, and fine arts.

The intercorrelations of the 80 questionnaire items were computed across the 456 subjects. Principal components of this factor analysis correlation matrix was then obtained (using unit weights in the main diagonals). The 10 largest factors, all with eigenvalues larger than 1, were then orthogonally rotated by the varimax method. Although there were more than 10 factors with eigenvalues greater than 1, it was decided to limit rotation to the largest 10. This decision was based on grounds of conceptual clarity since it was felt that only a small number of factors would be explanatory.

#### RESULTS

Of the 10 rotated factors, Factors 4, 5, 8, and 9 lacked conceptual focus and could not be interpreted. Factor loadings of selected items on the remaining six factors are shown below following each item.

#### *Description of Factors*

*Global Mechanism* (Factor 1) contains items that reveal a positive or negative *global* attitude toward technology. Included in this scale are items that indicate the stressful nature of technology [e.g., "Technological change is occurring so fast people are becoming second to machines" (.67)], items that express lack of confidence in technological cures [e.g., "In order to solve the problems of environmental pollution, mankind should stop using machines that pollute, rather than attempt to develop new machines that purportedly will be cleaner (.47)] as well as items that express a low valuation for the products of technology [e.g., "The greatest reason the dollar is worth so little today is that most goods are produced by machine (.54)]."

*Mechanical Curiosity* (Factor 2) contains items that express mechanical competence [e.g., "Computers are so foreign to me that I have little understanding of them (.36)]," as well as items that express curiosity for machines [e.g., "I have never had any desire to learn how a car engine operates (.55); "I would prefer reading *Popular Mechanics* to reading *Life* (-.61)]." Other items on this scale express a relative preference for technical rather than humanistic events [e.g., "I prefer building models to reading books (-.54) . . . "If I were in a recording studio, I would probably be more interested in the equipment used in making a record than in listening to the music (-.57)]."

*Preference for handmade goods* (Factor 3) is narrowly defined by items directly related to this concept [e.g., "A handmade gift is generally more appreciated than one which is mass produced (-.45) . . . "The only real quality items on the market are handmade (-.43)]."

*Alienation* (Factor 4) is composed of items that appear to reflect societal unconcern with the individual [e.g., "I feel I have no more meaning to the university than a pack of computer cards (-.43) . . . "Nowadays it is hard for one man to leave his mark on society (-.36)]."

*Spiritual Benefits of Technology* (Factor 5) contains items that consider technology as a

TABLE 1  
MEANS AND F RATIOS FOR VARIMAX FACTOR SCORES

Factor	Mean standardized score					Standard discriminant function coefficients
	Physical science	Biological science	Social science	Fine arts	F	
1. Global Mechanism	.10	.03	-.04	-.06	<1	-.15
2. Mechanical Curiosity	.62	.23	-.36	-.29	22.7*	-.94
3. Handmade Goods	-.34	-.00	.08	.10	2.9**	-.33
4. Alienation	-.23	.12	-.02	-.06	2.1***	.03
5. Spiritual Benefits	.02	-.08	.08	.00	<1	.12
6. Human Vitalism	.10	-.02	.03	-.11	<1	-.05

\* =  $p < .001$ .

\*\* =  $p < .05$ .

\*\*\* =  $p < 10$ .

"deus ex-machina," a rapid and dramatic way of solving more problems [e.g., "A true machine age will enable man to achieve the promise of a rich and rewarding spiritual life (.41) . . . "Increased mechanization will free mankind to engage in lofty pursuits (.50)]," as well as items that express the belief that technology can cure its own ills [e.g., "Halt ing technological change would be like confining ourselves to a permanent state of misery (.56)]."

*Human Vitalism* (Factor 6) contains items that reflect the belief that there is a "human element" which machines cannot duplicate [e.g., "Computers will never be able to think as creatively as man (.57) . . . "Poets and composers can contribute to understanding this world more than high speed computers can (.35)]." In addition, this scale seems to reflect the belief that much of modern technology would not be hard to master [e.g., "Any machine which is easy to operate and reliable is really very simple in conception (.38) . . . "It takes a far greater talent to write good poetry or prose than to design a complicated machine (.35)]."

#### Factor Scores

While the factor analysis is useful in reducing the complexity of the attitude domain, it was desirable to assess the discriminating power of the obtained dimensions in comparing criterion groups. To accomplish this scores for each factor were calculated using a method suggested by Glass and Maguire

(1966, p. 302). The matrix of factor scores  $X$  is derived as follows:

$$X = (F^T F)^{-1} F^T Z, \text{ where}$$

$Z$  is the  $n$  by  $N$  matrix of scores on variables,

$X$  is the  $m$  by  $N$  matrix of scores on factors, and

$F$  is the  $n$  by  $m$  matrix of "factor loadings" called the "factor pattern."

Factor scores created in this manner will be orthogonal. (An empirical check of the correlations between factor scores supported this statement insofar as all correlations were zero.)

#### Group Comparison on Factor Scores

The mean normalized factor scores for each group as well as the univariate  $F$  ratios for between-group comparisons are presented in Table 1. It can be seen from Table 1 that group differences on two factors are significant below the .05 level, one of which is significant well beyond the .001 level.

Multiple univariate comparisons are less informative than a single multivariate comparison since the dimensionality of group differences cannot be assessed through univariate methods. Thus, discriminant analysis (Rao, 1952) was performed, using the six factor scores as dependent variables.

The significance of the discriminating power of the six factors is indicated by a significant difference between group mean vectors using Rao's approximation of the  $F$

ratio ( $F [18,1264] = 4.85; p < .001$ ). Only the largest root of  $W^{-1}A$  (where  $W^{-1}$  = the inverse of the within groups dispersion matrix, and  $A$  = the among groups matrix) was significant ( $\chi^2 [18] = 85.02; p < .001$ ). This root accounted for approximately 87% of the canonical variation among groups. It, therefore, appears that the differences among the four groups can be represented along a single dimension. This dimension (discriminant function) is best defined by the relative weights of the factors that compose it. From Table 1 it can be seen that Factors 2 and 3 have the largest discriminant weights. Thus, it would appear that the differences between groups are largely defined by mechanical curiosity and, to a lesser extent, by a preference for handmade goods.

### DISCUSSION

It appears that independent dimensions of attitudes toward mechanization in this study do not all discriminate between students in different major fields. It is interesting to note that a global favorable or unfavorable attitude toward mechanization did not discriminate between the student groups used in this study. A tempting a priori assumption might state that science students would hold more favorable attitudes toward technology than nonscience students. Our results seemed to indicate that the *Mechanical Curiosity* factor was the major difference between science and nonscience groups. It appears that mechanical curiosity does not necessarily imply a value judgment concerning the outcomes of technology.

One of the implications of the present study is that defenders of technology might not be found exclusively in the ranks of its practitioners. Another interesting implication is that the choice of a major field is very strongly related to feelings of curiosity (very likely competence as well) about technology, a conclusion that certainly would be supported by research findings on interests. Future studies might be directed toward the further development of scales for the several meaningful factors identified.

### REFERENCES

- CALDER, N. (Ed.) *The complete new scientist series*. Vols. 1 & 2. Harmondsworth, England: Penguin Books, 1965.
- CAMPBELL, D., BORGES, F., EASTES, S., JOHANSSON, C., & PETERSON, R. A set of basic interest scales for the Strong Vocational Interest Blank for Men. *Journal of Applied Psychology Monographs*, 1968, 52, 1-54.
- GLASS, G., & MAGUIRE, T. Abuses of factor scores. *American Educational Research Journal*, 1966, 3, 297-304.
- KAHN, H., & WIENER, A. *The year 2000*. New York: Macmillan, 1967.
- KRECH, D. Controlling the mind controllers. *Think*, 1966, 32, 3-7.
- MCKECHNIE, G. Measuring environmental dispositions with the environmental response inventory. Paper presented at the 1970 Conference of the Environmental Design Research Association, Pittsburgh, October 28-30, 1970.
- PREHODA, R. *Designing the future: The role of technological forecasting*. Philadelphia: Chilton Books, 1967.
- RAO, C. *Advanced statistical methods in biometric research*. New York: Wiley, 1952.
- TOFFLER, A. *Future Shock*. New York: Random House, 1970.

(Received August 11, 1971)

## SHORT NOTES

### AN EVALUATION OF ITEM-BY-ITEM TEST ADMINISTRATION<sup>1</sup>

CECIL J. MULLINS<sup>2</sup> AND IRIS H. MASSEY

Air Force Human Resources Laboratory, Personnel Research Division,  
San Antonio, Texas

A battery of three tests was administered to two groups of basic airmen in their first week of basic training. Group A ( $N = 298$ ) was tested in the normal way; Group B ( $N = 317$ ) was tested with an item-by-item form of administration. The purpose was to determine whether the item-by-item administration would be more efficient than the usual method. Results did not indicate that the item-by-item administration was in any way superior to the usual.

Test anxiety and inability to read and understand test items could be factors influencing test results collected in the usual way. If these factors are important, the effect should be that scores routinely collected on a test designed to measure a particular single factor (i.e., mechanical aptitude), might be measuring effects not intended by the test designer. In addition to measuring mechanical aptitude, for example, the test may be measuring unrelated abilities, such as ability to follow directions, inability to work independently, reading speed and comprehension, and test anxiety. To the extent that these unrelated abilities are reflected by the subject's score, the test is no longer a single-factor test, and its validity against a particular criterion becomes uncertain. If the criterion is not composed in large measure of these same "unrelated factors," then the predictor test will be less valid than it might have been if given under unusual test conditions which minimize the influence of the unrelated abilities.

This study was devised to evaluate a method of administering tests, item by item, with the test administrator reading each item aloud and requiring all subjects to respond to that item before going on to the next one. This approach should decrease the effects of test anxiety and reading proficiency on the test, with the following consequent effects:

1. A decrease in intercorrelations among predictor variables collected by this administration method, since these scores should no longer con-

tain as much of the "unrelated" factors in common.

2. An increase in item-item reliability, since scores collected in this way should be less complex.

3. An increase in validity for any criterion measure which does not depend heavily on the "unrelated" factors eliminated or reduced by this method of administration.

### METHOD

A battery consisting of three tests—General Mechanics Test, Reading Comprehension Test, and Arithmetic Reasoning Test—was administered to a total of 629 basic airmen in their first week of training. The battery was administered under two different testing conditions on alternate days of testing. Group A ( $N = 298$ ) was tested in the normal way; Group B ( $N = 317$ ) was tested with the item-by-item form of administration. Careful time records were kept for both forms of administration to determine administrative feasibility of item-by-item administration if it proved to increase validity. The differences in testing time required for the two forms of administration were small. The average time required for item-by-item testing was 6 minutes longer for Reading Comprehension, 5 minutes shorter for Arithmetic Reasoning, and 16 minutes longer for General Mechanics.

Means, standard deviations, and intercorrelations of all variables were computed for Groups A and B combined ( $N = 615$ ) and for Group A ( $N = 298$ ) and Group B ( $N = 317$ ) separately. Comparison of the means and standard deviations shows no practical difference between Group A and Group B in educational background, Armed Forces Qualification Test (AFQT) score, or any of the Aptitude Indexes from the Airman Qualifying Examination (AQE).

Intercorrelations among the three experimental test scores did not differ consistently across the two kinds of administration. The expected differences between the two groups in odds-evens reliability coefficients, corrected by the Spearman-Brown prophecy formula, did not materialize. There were

<sup>1</sup> The research reported in this article was conducted by personnel of the Personnel Division, Air Force Human Resources Laboratory, AFSC, United States Air Force, Lackland AFB, Texas. Further reproduction is authorized to satisfy the needs of the U.S. Government. The views expressed here are those of the authors and do not necessarily reflect the views of the United States Air Force or the Department of Defense.

<sup>2</sup> Requests for reprints should be sent to Cecil J. Mullins, Selection Systems Section, Headquarters Personnel Research, Division AFHRL (AFSC), Lackland Air Force Base, Texas 78236.

no consistent differences in reliabilities between the two groups.<sup>2</sup>

The subjects were then cross-matched with a criterion file to collect pass-fail and final school grade criterion scores. Because of the matching, 97 cases were lost in the Normal Administration Group and 125 were lost in the Item-by-Item Administration Group. Intercorrelation matrices, along with means and standard deviations were again computed for the Total, Group A and Group B combined ( $N = 393$ ), and for Group A ( $N = 201$ ) and Group B ( $N = 192$ ) separately.

Since the sample had now shrunk to 393, there was no possibility of keeping criterion groups separate for the validation phase. The two criterion measures represented performance in many different schools. In some situations, this mixing together of different kinds of criterion score would cause severe experimental difficulties. In this particular study, however, it should make little difference. Validities for the total sample were undoubtedly depressed because the criteria were mixed, but it is the comparison of validities across Samples A and B that is of interest in this investigation, and there are no known biasing effects which could have operated in the composition of the two subsamples so that validities in one would have been artificially depressed relative to the other.

Comparison of validities of the three experimental tests for the two types of administration indicated no practical difference between the two sets against final school grade. Differences between the two groups in validities of the experimental predictors against the pass-fail criterion were universally in favor of the normal method of administration.

Finally, the data were subjected to a treatment similar to that described by Bottenberg and Christal (1961). Briefly, the technique calls for computation of an  $R^2$  between the criterion and the set of predictors for the total sample, and another similar  $R^2$  for each of the two subsamples. The  $R^2$ s for the separate treatment groups are then combined in such a way as to reveal what the  $R^2$  for the total sample would be if the weights for the predictors and the criterion means were free to vary optimally within each subsample. This combined  $R^2$  is then compared with the total  $R^2$  (which assigns the same

<sup>2</sup> Tables illustrating all referenced statistics are available from the authors on request (see address in Footnote 1).

weights for the predictor variables for each person in the study regardless of subsample, and which assumes the same criterion mean for all subjects). If the difference between the combined  $R^2$  and the total  $R^2$  is significant, the interpretation is that there are significant differences among the relative weights or criterion means from one treatment group to the other.

## RESULTS

There were no significant differences between the total  $R^2$  and the combined  $R^2$ , computed for the 3 experimental test scores, for either criterion (.236 and .237 against final school grade and .044 and .051 for pass/fail); or for similar  $R^2$  comparisons when the three tests were combined with Education, AFQT, and AQE scores to form a larger predictor set (.274 and .292 against final school grade and .068 and .101 against pass/fail).

It is conceivable that the item-by-item method of administration might affect the performance of low-ability airmen while leaving the performance of other subjects unchanged. Mean scores on the experimental tests were compared across groups A ( $N = 8$ ) and B ( $N = 14$ ) for all subjects in this study with less than 12 years of education, and again for all subjects who scored below 30 on the AFQT (Group A,  $N = 61$ ; Group B,  $N = 83$ ). There were no differences significant at the .05 level between the two groups. These were small groups, but there is no encouragement in the comparisons.

## DISCUSSION

Administering these three tests item by item, as opposed to the normal method of administration, did not result in any indication at all that this method of administration of tests was in any way superior to the older standard method.

## REFERENCE

- BOTTENBERG, R. A., & CHRISTAL, R. E. *An iterative technique for clustering criteria which retains optimum predictive efficiency*. (Tech. Rep. No. WADD-TH-61-30, AD-261 615) Lackland AFB, Tex.: Personnel Laboratory, Wright Air Development Division, March 1961.

(Received September 13, 1971)

## A LONGITUDINAL PREDICTIVE STUDY OF SUCCESS AND PERFORMANCE OF LAW ENFORCEMENT OFFICERS<sup>1</sup>

STANLEY P. AZEN<sup>2</sup>

*University of Southern California*

HOMA M. SNIBBE AND  
HUGH R. MONTGOMERY

*Occupational Health Service,  
Los Angeles County Department of Personnel*

A 20-year longitudinal study of biographical, psychological, and aptitudinal variables predictive of successful police performance is described. Subjects were 95 men appointed as deputy sheriffs in the Los Angeles County Sheriff's Department between 1947 and 1950. Among the significant predictors, stepwise-discriminant analysis yielded as "best" predictors of at least one criterion of success, age, height, the civil service written test score, scale 9 of the MMPI, the Kuder Mechanical scale, and the Guilford-Martin General Activity scale.

There is need for work identifying reliable and valid predictors of police performance (for a review, see Becker & Felkenes, 1968). Although most studies have involved concurrent validity (e.g., Sterne, 1960; Baehr, Furcon, & Froemel, 1968), a 7-year prediction study by Blum (1964) found correlations between predictors (including MMPI scales) and such criteria as supervisors' ratings, misconduct and commendations.

Marsh (1962), in a 10-year predictive study of sheriffs employed in Los Angeles County, established that certain performance criteria could be successfully predicted by various test scores, ratings and biodata. The present paper reports the results of a follow-up of some of these same officers after a 20-year period. The primary objectives of the study were (a) to evaluate the continued significance of Marsh's 10-year predictors as 20-year predictors and (b) to determine for each criterion the "best" among those predictors found significant.

### METHOD

#### Subjects

The subjects were 95 male law enforcement officers chosen from two randomly selected classes from the Los Angeles County Sheriff's Academy. All subjects were appointed as deputies in 1947-1950.

#### Variables

Variables found predictive of job performance in Marsh's study are: height (inches); age (years); written test scores on civil service examination (standard score based on general ability, practical judgment and memory); score on General Activity scale on Guilford-Martin Temperament Inventory (Guilford & Martin, 1934); scales 9 (Hypomania), 1 (Hypochondriasis) and 2 (Depression) on MMPI; Mechanical and Social Service scores on Kuder Vocational Preference Record; and rating<sup>a</sup> at Sheriff's Academy.

Six dichotomous criterion measures of success and performance of subjects up to 1970 or prior to termination were studied. The criteria and criterion groups were: employment status as of 1970 (employed or not); rank status as of 1970 (promoted or not); job type as of 1970 or termination date (patrol or other); average of all supervisors' ratings (low or high); job related auto accidents prior to 1958 (none or at least one); and job related auto accidents prior to 1970 (none or at least one).

These criteria are not identical to those used by Marsh. For practical reasons, we omitted Marsh's special, forced-choice supervisors' ratings, substituting the average of all supervisors' routine, semi-annual ratings from time of appointment to termination or 1970. Marsh included only patrol-car accidents judged to be "nonpreventable," while our data did not distinguish between "preventible" and "nonpreventable." Finally, we added rank status and job type, and incorporated Marsh's criterion of discharge into employment status.

#### Procedure

One-way analysis of variance was performed to determine the significant predictors for each criterion. For each criterion, the significant predictor with the largest *F* was selected as the "best" pre-

<sup>1</sup> Work on this study was supported by LEAA Grant No. 72-DF-09-0005, administered under the auspices of the Los Angeles County Sheriff's Department and the Department of Personnel. The Physiological Fitness Standards Research Project Staff Psychologists are J. Gencik, H. Snibbe, and H. Montgomery. The authors would like to express deep appreciation to the Graduate Assistant, Miss Lynn Maruyama, and to Mrs. Betty Horvath whose energy has contributed greatly to these efforts.

<sup>2</sup> Requests for reprints should be sent to Stanley P. Azen, School of Engineering, University of Southern California, University Park, Los Angeles, California 90007.

<sup>a</sup> Since data was available for only 45 officers, these data were analyzed independently.

TABLE 1

CRITERIA, THEIR PREDICTORS, AND PROBABILITIES OF CORRECT CLASSIFICATION

Criterion	Single and second best predictors	Probability of correct classification
Employment status	General Activity	0.59
Rank status	Civil Service Written Test Kuder Mechanical	0.66 0.72
Job type	Age Kuder Mechanical	0.60 0.63
Average supervisors' rating	Kuder Mechanical	0.67
Auto accidents 1947-1958	MMPI Scale 9 (Hypomania) Height	0.60 0.63
Auto accidents 1947-1970	MMPI Scale 9 (Hypomania)	0.64

dictor. A discriminant analysis was executed for each best predictor to derive a classification decision rule and an estimated probability of correct classification. A stepwise-discriminant analysis (Afifi & Azen, 1972) was then performed to identify a second "best" predictor given the single best predictor for each criterion. Classification rules and estimated probabilities of correct classification were also calculated for the second best predictors.

## RESULTS

From the analysis of variance ( $\alpha = .10$ ), the significant predictors for each criterion were: *employment status*—General Activity score ( $F = 5.11$ ,  $p < .025$ ); *rank status*—civil service written test score ( $F = 6.71$ ,  $p < .025$ ), Mechanical score ( $F = 6.44$ ,  $p < .025$ ) and MMPI scale 1 (Hypochondriasis) ( $F = 4.05$ ,  $p < .025$ ); *job type*—age ( $F = 5.62$ ,  $p < .025$ ) and Mechanical score ( $F = 2.89$ ,  $p < .10$ ); *supervisors' ratings*—Mechanical score ( $F = 5.57$ ,  $p < .025$ ); *auto accidents 1947-1958*—MMPI scales 9 (Hypomania) ( $F = 4.61$ ,  $p < .05$ ) and 2 (Depression) ( $F = 2.76$ ,  $p < .10$ ) and height ( $F = 2.75$ ,  $p < .10$ ); *auto accidents 1947-1970*—MMPI scales 9 (Hypomania) ( $F = 9.70$ ,  $p < .005$ ) and 2 (Depression) ( $F = 3.69$ ,  $p < .10$ ). Social Service score and Academy rating had no significant relationship to any criterion.

Table 1 contains a summary by criterion, of the "best" and second "best" predictors from the stepwise discriminant analysis. The "probability of correct classification" specifies the likelihood that a subject will be assigned to a given criterion group.<sup>4</sup>

<sup>4</sup> Space limitations preclude presentation of classification decision rules. These are available from the senior author (see address in Footnote 2).

## DISCUSSION

Three central results emerge from the analysis of variance: (a) some significant predictors over 20 years were found, (b) some of these agree with Marsh's work (despite differences in criteria, sampling and analysis), and (c) different criteria are predicted by different predictors.

The significant predictors over 20 years are listed above; those directly supporting Marsh's earlier work are the MMPI predictors of auto accidents. That MMPI scales 9 (Hypomania) (directly) and 2 (Depression) (inversely) are related to auto accidents over the first and second 10 years of police career is a remarkable finding, which may have implication for non-police drivers, also.

The principal result of the stepwise-discriminant analysis is that the Kuder Mechanical score emerges as the most generally useful predictor of the criteria (since it predicts 3 of the 6 criteria). This is also an exciting finding since a) one problem with validation studies is the failure of predictors to relate significantly to more than one criteria, and b) the nature of this result is anticipated in the work of Cattell, Eber & Tatsooka (1970).

One problem with these results is generic to longitudinal studies: some of the test predictors have undergone revision and the forms demonstrated to have validity may not be generally available. The Guilford-Martin, for example, is now the Guilford-Zimmerman Temperament Survey (1949). The civil service test, created by the Personnel Department of Los Angeles County, has undergone numerous revisions. Thus, the utility of such results is attenuated. Additionally, the role of the police officer and the definition of "success" has changed a great deal (Silver, 1967). Such changes can be expected to continue at an increasingly rapid rate, pointing to the need for continual and rigorous selection research.

## REFERENCES

- AFIFI, A. A., & AZEN, S. P. *Statistical analysis: a computer oriented approach*. New York: Academic Press, 1972.
- BAEHR, M. E., FURCON, J. E., & FROEMEL, E. C. *Psychological assessment of patrolman qualifications in relation to field performance*. Washington, D.C.: Law Enforcement Assistance Administration, U.S. Department of Justice, 1968.
- BECKER, H. K., & FELKENES, G. T. *Law enforcement: A selected bibliography*. Metuchen, New Jersey: Scarecrow Press, 1968.
- BLUM, R. H. *Police selection*. Springfield, Ill.: Charles C Thomas, 1964.
- CATTELL, R. B., EBEL, H. W., & TATSUOKA, M. M. *Handbook for the sixteen personality factor questionnaire*. Champaign, Ill.: Institute for Personality and Ability Testing, 1970.
- GUILFORD, J. P., & MARTIN, H. G. *Guilford-Martin Temperament Profile charts*. Chicago: Sheridan Supply Co., 1934.

- GUILFORD, J. P., & ZIMMERMAN, W. S. *Guilford-Zimmerman Temperament Survey*. Chicago: Sheridan Supply Co., 1949.
- MARSH, S. H. Validating the selection of deputy sheriffs. *Public Personnel Review*, 1962, 23, 41-44.
- SILVER, A. The demand for order in civil society: A review of some themes in the history of urban crime, police, and

riot. In D. J. Bondua (Ed.), *The police*. New York: Wiley, 1967.

STERNE, D. M. Use of the Kuder Preference Record-Personal with police officers. *Journal of Applied Psychology*, 1960, 44, 323-324.

(Received for early publication, March 31, 1972)

*Journal of Applied Psychology*  
1973, Vol. 57, No. 2, 192-194

## DIMENSIONAL ANALYSIS OF THE LEAST PREFERRED CO-WORKER SCALES<sup>1</sup>

WILLIAM M. FOX,<sup>2</sup> WALTER A. HILL, AND WILSON H. GUERTIN

*University of Florida*

This article presents comparative factor analyses of responses to Fiedler's least preferred co-worker scales based on the responses of three samples. It appears possible to identify several dimensions of coworker perceptions measured by the LPC scales.

The contingency model of leadership effectiveness proposed by Fiedler (1967) and his associates predicts that managers scoring low on the least preferred co-worker (LPC) questionnaire will be more effective when the situation is either very favorable or very unfavorable to exert influence and that supervisors scoring high on this instrument will be more effective in situations characterized by intermediate favorability. The LPC, a leadership style score that is thought to measure one's esteem for his least preferred co-worker, is obtained by asking an individual to think of everyone with whom he has ever worked and then to describe the person with whom he could work least well on a series of bipolar scales such as

friendly:	:	:	:	:	:	:	unfriendly
8	7	6	5	4	3	2	1
efficient:	:	:	:	:	:	:	inefficient
8	7	6	5	4	3	2	1

This instrument has been found effective in predicting leadership effectiveness in many diverse situations (Blanchard, 1967; Hill, 1969; Hopfe, 1970; Hunt, 1967) despite the fact that the precise meaning of LPC has not been determined. Fiedler (1970) indicates that the LPC is still uncorrelated with most personality and cultural scores and various attempts to relate it

to self descriptions, descriptions by others, or to behavioral observations have led to complex or inconsistent results.

The purpose of this article is to attempt to develop a better conceptual understanding of the Least Preferred Co-worker instrument. This will be accomplished by factor analyzing the responses of subjects from two different organizations in the United States and from a sample of English managers.

### METHOD

#### Measures

The original 16-item LPC instrument (Fiedler, 1967) was administered to Internal Revenue Service tax examiners and to English managers. Additionally, an LPC, consisting of 24 items, was administered to a sample of U.S. Marines; the first 16 items in this form were the same as those in the original LPC instrument. In selecting additional items for the revised instrument, an attempt was made to select items which were not unduly redundant with the original 16 items, as well as items which were characteristic of people one finds difficult to work with (Fox, Hill, & Guertin, 1971).

#### Subjects

The sample of 114 Internal Revenue tax examiners consisted of supervisors, assistant supervisors, and clerical personnel whose main function was to audit income tax returns.<sup>3</sup> All respondents were located at a regional headquarters of the IRS. The U.S. Marine sample of 147 consisted of squad leaders, fire team

<sup>1</sup> This article was prepared in connection with research done under the Office of Naval Research, Group Psychology Programs, Contract No. N00014-68-A-0173 0010.

<sup>2</sup> Requests for reprints should be sent to William M. Fox, College of Business Administration, University of Florida, Gainesville, Florida, 32601.

<sup>3</sup> The questionnaires were actually administered twice, four weeks apart, and all 228 questionnaires were used in the analysis. The effects on the correlations are unknown.

TABLE 1  
COMPARISON OF LPC FACTOR ANALYSES OF THREE SAMPLES

Items	1			2			3			4			5		
	Hostile-Ineffective			Remote-Rejecting			Tense			Boring-Ineffective			Hesitant		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Unpleasant				.70	b	.74									
Unfriendly	.16	.27	.53	.81	b	.18									
Rejecting	.57	.21	.04	.38	.64	.72									
Frustrating	.48	.76	.06												
Unenthusiastic															
Tense							-.13	.03	.70						
Distant							.62	.61	.66						
Cold															
Uncooperative	.77	.73	.10	.77	.57	.84									
Hostile	.74	.54	.70	.30	.22	.66									
Boring															
Quarrelsome	.67	.30	.60	.23	.51	.08									
Hesitant															
Inefficient	.38	.66	.67												
Gloomy	.18	.30	.21	.56	.30	.39									
Guarded	.20	.17	.60												

\* A = IRS subjects; B = Marine subjects; C = English managers.

<sup>b</sup> The rotated matrix for Marines split this factor into two so that these two loadings are in the .70's on a factor not reported here. If factor space is unduly compressed the two factors coalesce.

leaders, and squad members of two companies of a training battalion located in the Southeastern portion of the United States. The sample of 180 English managers was drawn from three separate electronics firms located in the Midlands, England. Although almost all of these managers were from research and development, a few accounting managers were included.

### Analysis

The Guertin and Bailey (1970) library of factor analytic programs provided the Varimax and Simple Loadings (primary) factor rotations. Iterated communalities were used in the diagonals of the correlation matrix to give principal axes. In all analyses multiple trial rotations were made with different numbers of principal axes. Final selection of the number of factors rotated was based upon considerations summarized in Guertin and Bailey (1970, p. 121). The orthogonal solutions reported here were only employed after the oblique Simple Loadings solutions failed to clarify factor structure to an appreciably better degree.

### RESULTS

Since complete matrices derived from the data are available elsewhere (Fox, Hill, & Guertin, 1971), only the most relevant summary findings will be given here. Emphasis will be on similarities in factor structure across samples so attention can be focused on the more stable factors underlying the responses to LPC items.

The IRS data produced the simplest factor space with only four dimensions. The addition of the fifth principal axis only accounted for another 2.7% of the score variance. The Marine

and English data both required six factors to represent common factor space adequately, although only five will be discussed. Between 54.9 and 62.5% of total variance was explained by these three Varimax rotated matrices.

A simplified composite of the three Varimax matrices is given in Table 1. Only items with loadings of more than .50 in at least one sample are included. For these items, corresponding loadings for the other two samples are shown even though, in some cases, they were small.

The first factor is called *Hostile-Ineffective*. Factor loadings hold up best across samples for the "hostile" item. The "quarrelsome," "frustrating" and "inefficient" items add additional substance to the factor. This factor underlying the LPC description would make the co-worker appear to be a person who openly expresses great felt hostility by being uncooperative and quarrelsome. Quite naturally, the failure to be able to work together would lead to an ineffective effort.

The second factor is called *Remote-Rejecting*. Factor loadings seem to hold up best across samples on the "distant" and "gloomy" items. As the footnote reference to entry b in Table 1 explains, the sixth factor of the Marine data pulled the loadings for the first two LPC items away from this second factor. The U.S. sample associates the items "unpleasant" and "unfriendly" while the English sample does not. The person described on the LPC in terms of this factor would seem to be unpleasant because of remoteness, a gloominess and self-concern. He would

reject a cooperative relationship and be hard to get to know.

The third factor cross-validates only on the "tense" item, so it is called simply *Tense*. The only other item that comes into the picture is an essential "coldness." The English sample identifies lack of enthusiasm with this tense-coldness but neither U.S. sample does.

The fourth and last factor found in the IRS responses is called *Ineffective-Boring*. The English sample failed to cross-validate the configuration of items that made up the factor for the IRS sample. Actually only the "boring" item is strongly loaded in the Marine analysis. The English would seem to use the attribute "boring" in a somewhat different way than do the Americans. The "inefficient" item is not strongly loaded, but common to both American analyses. Thus, the factor seems to underlie the description of a boring co-worker who is inefficient and ineffective in getting work done by team effort.

The fifth and last factor to be described appeared in the Marine and English data. It is called *Hesitant*. The items showing loadings are "hesitant" and "unenthusiastic." The suggestion is one of a "burnt-child" reaction. The English sample associates the "gloomy" item with this factor but the U.S. sample does not.

#### DISCUSSION AND CONCLUSIONS

Two possible shortcomings in this analysis should be pointed out. On the one hand, the English sample consisted solely of managers while the U.S. samples involved both managers and subordinates. The small number of managers in the U.S. samples precluded a separate analysis of managers. On the other hand, a well recognized cultural language difference exists between English and American respondents. This difference may result in different interpretations being

given to the same items. Both of these shortcomings may cloud the comparison between the U.S. and U.K. subjects.<sup>4</sup> Although this study is preliminary, it does appear that the LPC measures several identifiable components of co-worker perceptions. Pending more definitive analyses, these appear to reflect perceptions of least preferred co-workers in terms of "hostile-ineffective," "remote-rejecting," "tense," and "hesitant" dimensions. Further research should test these hypotheses and evaluate if scores on these different dimensions are differentially related to leader-group effectiveness.

<sup>4</sup> Since it is extremely difficult for an American to interpret not only items but also factors that evolve in a cross cultural study, we asked three English nationals to assist by independently naming the factors which emerged from the English data. Among the differences between our labels and those of the English nationals, one was most outstanding. All three English nationals labeled the elements gloomy, hesitant and unenthusiastic as "rigid." This may indicate a conceptual as well as a semantic difference.

#### REFERENCES

- BLANCHARD, K. H. College boards of trustees: A need for directive leadership. *Academy of Management Journal*, 1967, **10**, 409-417.
- FIEDLER, F. E. *A Theory of leadership effectiveness*. New York: McGraw-Hill, 1967.
- FIEDLER, F. E. Personality, motivational systems, and behavior of high and low LPC persons. (Tech. Rep. No. 70-12) Seattle: Department of Psychology, University of Washington, 1970.
- FOX, W. M., HILL, W. A., & GUERTIN, W. H. Factor analysis of Least Preferred Co-worker Scales. (Tech. Rep. No. 70-1) Washington, D.C.: Office of Naval Research, 1971.
- GUERTIN, W. H., & BAILEY, J. P. *Introduction to modern factor analysis*. Ann Arbor: Edwards, 1970.
- HILL, W. The validation and extension of Fiedler's theory of leadership effectiveness. *Academy of Management Journal*, 1969, **12**, 33-47.
- HOPPE, M. W. Leadership style and effectiveness of department chairman in business administration. *Academy of Management Journal*, 1970, **18**, 301-310.
- HUNT, J. G. Fiedler's leadership contingency model: An empirical test in three organizations. *Organizational Behavior and Human Performance*, 1969, **2**, 290-308.

(Received October 18, 1971)

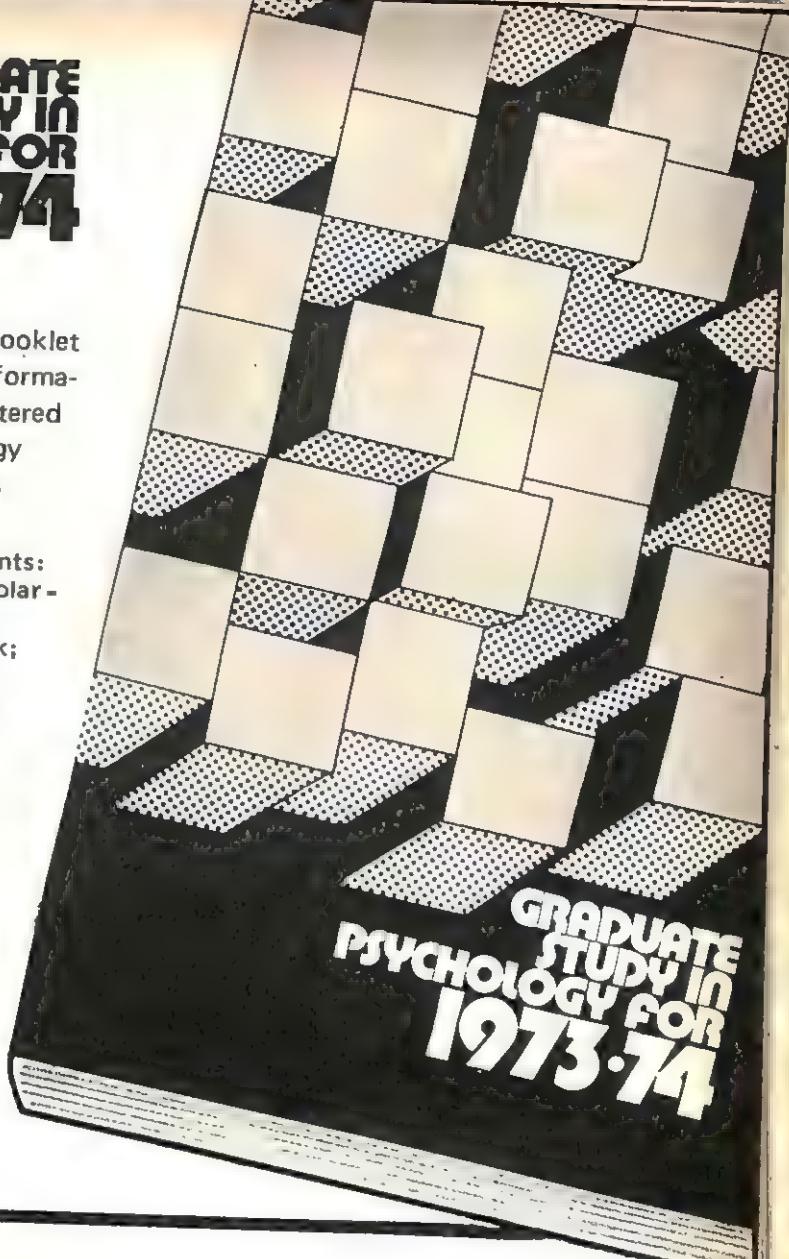
# **GRADUATE STUDY IN PSYCHOLOGY FOR 1973-74**

The new sixth edition of this booklet provides relevant up-to-date information on 395 separately administered graduate programs in psychology in 315 universities and colleges.

Areas covered include enrollments; application for fellowships, scholarships, or assistantships; types of assistantships and hours of work; government stipends; and post-doctoral arrangements. Current information has been supplied by the department chairman, his or her representative, or the program director.

**\$2** per  
copy

All orders amounting  
to \$15 or less  
must be prepaid.



total enclosed for

copies of **GRADUATE STUDY IN PSYCHOLOGY FOR 1973-74**

Order Department  
AMERICAN PSYCHOLOGICAL ASSOCIATION  
1200 Seventeenth Street, N.W.  
Washington, D.C. 20036

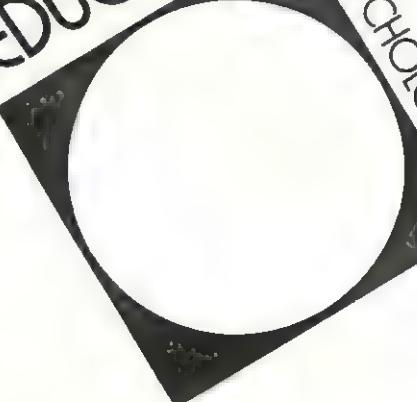
Name \_\_\_\_\_

Address \_\_\_\_\_

Zip \_\_\_\_\_

*newly released*

# UNDERGRADUATE EDUCATION IN PSYCHOLOGY



*Diversity in ideology, approach, method, and even epistemology is revealed in this survey of undergraduate education in psychology as it entered the 1970's. Reporting in-depth on programs at 10 institutions selected to reflect the range of philosophies and innovative teaching techniques in undergraduate psychology, this volume discusses major curriculum trends in both 2- and 4-year institutions.*

*Sponsored by APA, the survey was conducted by the Center for Research on Learning and Teaching at the University of Michigan. Primary authors: James A. Kulik, with Donald R. Brown, Richard E. Vestewig, and Janet Wright. To obtain copies, write to Order Department W, American Psychological Association, 1200 17th St., N.W., Washington, D.C. 20036. Prepayment of \$2 per copy must accompany all orders.*

# JOURNAL OF APPLIED PSYCHOLOGY

Copyright © 1973 by the American Psychological Association, Inc.

June 1973

Vol. 57, No. 3

## ARTICLES

- 708 (2)  
B 10 4 11 174  
ble
- |   |  |     |
|---|--|-----|
| The Characteristics of Subject Matter in Different Academic Areas   | Anthony Biglan   | 195 |
| Relationships between Subject Matter Characteristics and the Structure and Output of University Departments | Anthony Biglan   | 204 |
| Statistical Accuracy and Practical Utility in the Use of Moderator Variables                                | Craig C. Pinder  | 214 |
| A Method for Evaluating Alternative Recruiting-Selection Strategies: The CAPER Model                        | William A. Sands   | 222 |
| Response Requirements and Primacy-Recency Effects in a Simulated Selection Interview                        | James L. Farr  | 228 |
| Training Interviewers to Eliminate Contrast Effects in Employment Interviews                                | Kenneth N. Wexley, Raymond E. Sanders, and Gary A. Yutk        | 233 |
| Effect of Race on Peer Ratings in an Industrial Situation   | Frank L. Schmidt and Raymond H. Johnson                        | 237 |
| An Approach for Determining Criteria of Sales Performance   | David W. Cravens and Robert B. Woodruff                        | 242 |
| The Perception of Organizational Climate: The Customer's View   | Benjamin Schneider   | 248 |
| Performance Effectiveness and Efficiency Under Different Dyadic Work Strategies                             | Samuel C. Shiflett   | 257 |
| Experimental Test of the Valence-Instrumentality Relationship in Job Performance                            | Robert D. Pritchard and Philip J. De Leo                       | 264 |
| Effects of the Manipulation of a Performance-Reward Contingency on Behavior in a Simulated Work Setting     | Dale O. Jorgenson, Marvin D. Dunnette, and Robert D. Pritchard | 271 |
| Predicting the Emergence of Leaders Using Fiedler's Contingency Model of Leadership Effectiveness           | Robert W. Rice and Martin M. Chemers                           | 281 |
| Some Interactions between Personality Variables and Management Styles                                       | Kenneth E. Runyon  | 288 |
| Job Satisfaction among Whites and Nonwhites: A Cross-Cultural Approach                                      | Charles A. O'Reilly, III, and Karlene H. Roberts               | 295 |
| The Nature of Bias in Official Accident and Violation Records   | Frederick L. McGuire   | 300 |
| Prediction of Accidents in a Standardized Home Environment  | Joan S. Guilford   | 306 |
| The Recognition of Road Pavement Messages   | Wendy A. Macdonald and Errol R. Hoffmann                       | 314 |
| The Dynamic Role of Eye-Head Angular Displacements in Human Vehicular Guidance                              | Henry S. R. Kao  | 320 |
| Evaluating Language Translations: Experiments on Three Assessment Methods                                   | H. Wallace Sinaiko and Richard W. Brislin                      | 328 |
| The Relationship between Consumers' Category Width and Trial of New Products                                | James H. Donnelly, Jr., Michael J. Etzel, and Scott Roeth      | 335 |

## SHORT NOTES

- Career Orientation and Job Satisfaction among Working Wives  
*Martin J. Gannon and D. Hunt Hendrickson* 339
- Job Attitudes as Predictors of Termination and Absenteeism: Consistency over Time and  
across Organizational Units  
*L. K. Waters and Darrell Roach* 341
- Interview Decisions as Determined by Competency and Attitude Similarity  
*Glen D. Baskett* 343
- Effects of Sponsor and Prepayment on Compliance with a Mailed Request  
*Anthony N. Doob, Jonathan L. Freedman, and J. Merrill Carlsmith* 346
- Effects of Signed and Unsigned Questionnaires for both Sensitive and Nonsensitive Items  
*Richard P. Butler* 348  
*Richard L. Cahoon* 350
- Auditory Vigilance under Hypoxia  
*Dirk C. Prather* 353
- Prompted Mental Practice as A Flight Simulator  
*Benson Rosen, Thomas H. Jerdee, and W. Harvey Hegarty* 355
- Neuroticism among Policemen: An Examination of Police Personality  
*C. Abraham Fenster and Bernard Locke* 358
- Does Farm Practice Adoption Involve a General Trait?  
*James M. Richards, Jr., and John G. Clandy* 360  
338

## LIST OF MANUSCRIPTS ACCEPTED

This is the last issue of Volume 57.  
Volume Title Page and Contents appear herein.  
Information for Contributors appears on the last page of the Volume Contents.

## THE CHARACTERISTICS OF SUBJECT MATTER IN DIFFERENT ACADEMIC AREAS<sup>1</sup>

ANTHONY BIGLAN<sup>2</sup>

*University of Washington*

Multidimensional scaling was performed on scholars' judgments about the similarities of the subject matter of different academic areas. One hundred sixty-eight scholars at the University of Illinois made judgments about 36 areas, and 54 scholars at a small western college judged similarities among 30 areas. The method of sorting (Miller, 1969) was used in collecting data. Three dimensions were common to the solutions of both samples: (a) existence of a paradigm, (b) concern with application, and (c) concern with life systems. It appears that these dimensions are general to the subject matter of most academic institutions.

One of the most easily overlooked facts about university organization is that academic departments are organized according to subject matter. Typically, each field of specialization has its own department, and the department in which there is more than one discipline is the exception. Presumably this system arises from the peculiar requirements that each area has for the organization of its research, teaching, and administrative activities. While the organization of university departments has received increasing attention from social scientists (Menzel, 1962; Oncken, 1971; Pelz & Andrews, 1966), the way in which subject matter characteristics may require particular forms of department organization has not been examined. The chief reason for this is probably that there has not been a systematic analysis of subject matter characteristics that could serve as a framework for such a study. It is obvious that such fields as physics and psychology differ in subject matter, but what is the nature of these differences? This article presents a multidimensional analysis of this problem. A subsequent article to be presented in this journal (Biglan, 1973) uses the analysis of this study to examine relationships between subject matter characteristics and department organization.

<sup>1</sup> Research for this article was supported in part by the Office of the Executive Vice President and Provost, University of Illinois, Urbana, Illinois, and by the Department of Health, Education, and Welfare, Office of Education, Grant G-70-3347 (Fred E. Fiedler, principal investigator).

<sup>2</sup> Request for reprints should be sent to Anthony Biglan, Department of Psychiatry, University of Wisconsin, 427 Lorch Street, Madison, Wisconsin 53706.

How can we get at the "important" characteristics or dimensions of academic subject matter? In this study it was assumed that scholars in the various areas are the best source of information about the characteristics of different areas; whatever dimensions they use in thinking about academic areas are considered to be important and worthy of further investigation. Nonmetric multidimensional scaling (Kruskal, 1964a, 1964b; Shepard, 1962) provides an ideal method for determining these dimensions. The method employs subjects' judgments about the similarities (or differences) among a set of stimulus objects. From this ordinal data, a map or array of the stimulus points is developed in a metric multidimensional space that "best fits" the original data about the similarity of stimuli. In this way the technique provides metric scaling of the stimuli and, at the same time, indicates the dimensions that underlie subjects' perceptions of them. The technique allows comparison among all academic areas within the same framework but does not restrict the analysis to the oversimplification associated with a single dimension.

At least two dimensions are likely to be used by scholars when they think about academic subject matter. First, Kuhn has argued that the physical sciences are characterized by the existence of paradigms that specify the appropriate problems for study and the appropriate methods to be used. It appears that the social sciences and nonscience areas such as history do not have such clearly delineated paradigms. If this is true, we should find a dimension that distinguishes paradigm-

matic and nonparadigmatic fields. A second way in which scholars may perceive an area is in terms of its requirements for practical application. Thus, areas such as engineering and education are likely to be distinguished from areas such as English and chemistry.

### METHOD

Multidimensional scaling of subject matter characteristics was first performed on data obtained from scholars at the University of Illinois. Since the dimensions obtained in this setting could simply reflect the way areas are organized at large, state-supported universities, the scaling was replicated at a small, denominational liberal arts college in the State of Washington. If the same dimensions are used by scholars at both of these institutions, then we can be more certain that we are getting at characteristics of academic areas that are general and important. In addition, semantic differential ratings of each area on each of six attributes were obtained from scholars at the small college as an aid to interpreting the scaling results.

*Scaling technique.* Kruskal's (1964a, 1964b) technique for nonmetric multidimensional scaling was used in the present study. Nonmetric multidimensional scaling employs ordinal data about the similarity among a set of stimulus objects and generates a configuration of points in an  $n$ -dimensional metric space, such that the distances among points in the metric space maximally correspond to the ordinal similarity data. The number of dimensions,  $n$ , is specified by the user. The scaling begins with a random  $n$ -dimensional configuration. In an iterative procedure this configuration is changed in small steps in order to maximize its fit with the similarity data. Kruskal's measure of fit is called "stress." It ranges from 0 to 100%. Typically, solutions are generated for different values of  $n$ , and one solution is chosen as "best" on the basis of its stress value and the interpretability of its dimensions.

*The areas.* Thirty-six areas were included in the Illinois scaling. Included were such areas as Agricultural Engineering, Physics, and Philosophy. The areas were chosen to include as diverse a sample as possible. The availability of structure and output data was also considered in choosing areas. In the small college replication, all of the areas in which the college offered courses were included for scaling. In addition, four areas that had been used in the Illinois scaling were also used in the replication in order to allow comparison of the results of the two analyses.

*Judges.* One hundred and sixty-eight faculty members at the University of Illinois served as judges of area similarity. They were distributed over the 36 areas of interest with no more than five and no less than three judges in any area. Whenever possible, judges within an area were distributed over academic rank and subdisciplines. Only six faculty members refused to participate in the study when asked.

All of the approximately 70 faculty members at the small liberal arts college were asked to make judgments about the similarity of academic areas. They were contacted through the Dean of the College, who wrote

letters supporting the project. After one telephone follow-up by the Dean's office, 56 faculty members had returned completed judgments of which 54 were usable.

### Procedure

Most methods of collecting similarities data require judges to rate or rank the similarity of all pairs of stimuli. In the case of the Illinois scaling, such methods would require  $36(35)/2$  or 630 responses from each judge. Since it did not appear that university faculty could be prevailed upon to this extent, a procedure requiring fewer responses of each judge was needed. Such a procedure has been proposed by Miller (1969) and was used in the present study. The method of sorting required judges to put areas into categories on the basis of their similarity. No limit was placed on the number of categories. The judgements of one subject about the similarities among areas may be represented in an  $N \times N$  matrix whose rows and columns correspond to the academic areas of interest. Ones are placed in the cells of this matrix corresponding to the pairs of areas that were placed in the same category. Zeroes in cells indicate areas that were not placed in the same category. Summing over all judges' matrices provides a matrix whose cells indicate the number of judges who placed the pair of areas in the same category. Rao and Katz (1970) simulated the collection of similarities data using the sorting method. They compared the configuration obtained by scaling these similarities data with the known configuration they had started with. The correlation between the interpoint distances of the known configuration and the interpoint distance of the configuration obtained through the method of sorting was .81. This result compared favorably with the ability of other more common methods of collecting similarities data to recover the known configuration. Richards (in press) used real subjects in comparing the sorting method with a more common method of collecting similarities data. Canonical correlations between five-dimensional solutions for each method were .98, .96, .90, .60, and .46.

In collecting data at the University of Illinois, scholars were provided with  $36 \times 3 \times 5$  cards, each of which contained the name of one academic area. They were instructed to sort the cards into categories or piles on the basis of the similarity of the subject matter of each area. Data was typically collected in the scholar's office. Data from the small college replication were collected through the mails, using essentially the same procedure. In this case, the names of areas were presented on thirty slips of paper, and judges were asked to staple together the slips which they placed in the same category. Only one respondent appeared not to have understood these instructions. Upon completing the sorting task, scholars at the small college were asked to rate each area they had judged on the following bipolar adjectives: (a) pure-applied, (b) physical-nonphysical, (c) biological-nonbiological, (d) of interest to me personally of little or no interest to me personally, (e) traditional-nontraditional, and (f) life science-nonline science. Forms for these ratings were provided in a separate sealed envelope that judges were asked to leave sealed until they had completed the sorting task.

## RESULTS

*Scaling of the Illinois Data*

Kruskal's (1964b) MDSCAL program (Version 4M) was used to scale the area similarity data obtained from both samples. For the Illinois sample, solutions were obtained in six, five, four, three, and two dimensions. Kruskal's index of goodness of fit between the similarity data and the multidimensional solution is called stress. The stress values for these solutions were .078, .101, .127, .226, and .311, respectively. Each solution was rotated to principal axes in order to aid interpretation.

The three-dimensional solution was chosen as the "best" solution, since all three of its dimensions were interpretable and its stress value was .23. Kruskal's suggested verbal evaluation for this stress value is "fair." He adds, however, that "where data values are heavily replicated, this evaluation is pessimistic, and larger stress values are acceptable [p. 9]." Since there were 168 replications in the Illinois scaling, Kruskal's comment appears applicable.

The reliability of this configuration was evaluated by splitting the sample of judges into halves, obtaining a separate configuration for each half, and comparing these configurations. The judgments of all scholars who were in the first eighteen areas on an alphabetical list were placed in the first sample, and the remaining judgments comprised the second sample. A three-dimensional solution was obtained from the similarity judgments of each sample. The two configurations were compared by correlating the distances among each possible pair of stimuli in one configuration with the corresponding distances in the other configuration. This correlation was .88 ( $N = 630$ ). Thus, it appears that in the present circumstances the sorting method of data collection yielded stable results.

There is a second way in which the method of data collection used in the present study may yield unreliable configurations. Stimuli

may cluster rather than be evenly dispersed along the dimensions. This is not bad in itself, but with the data collection method used here the distances between points in different clusters may be less reliable than the distances between points in the same cluster. Visual inspection of the final three-dimensional solution from the Illinois sample did reveal clustering of areas. The areas could be grouped into eight clusters on the basis of their interpoint distances and visual inspection of the configuration. In order to test the reliability of intercluster distances, the two three-dimensional configurations described in the preceding paragraph were used. In both configurations, centroids were computed for each of the eight clusters of areas. The distances among the centroids in each configuration were then obtained. If intercluster distances are reliable, then there should be a high correlation between corresponding distances in the two configurations. This was, in fact, the case; the correlation was .88 ( $N = 28$ ). Thus, although clustering of stimuli occurred, it appears that the intercluster distances are reliable.

A third problem associated with the method of sorting is that individual differences in the perceptions of areas cannot be evaluated in the usual ways (cf. Carroll & Chang, 1969). Since the areas were clustered in eight sets in the accepted solution, one method of evaluating agreement among judges would be to compare the eight separate three-dimensional solutions that could be obtained from judgments of scholars in each of the eight clusters. These solutions were obtained and interpoint distances in each solution were correlated with the interpoint distances in every other solution. The correlations ranged from .61 to .84. The average was .75. No configuration stood out as different from the rest. These results suggest that faculty members in our sample perceive the relationships among areas in substantially the same way, regardless of their own area.

Figures 1, 2, and 3 present plots of the three-dimensional solution. Each dimension is plotted against the other two so that there are three two-dimensional plots. In Figure 1, dimension one is plotted along the horizontal dimension, and dimension two appears vertically.

On the first dimension, physical science and engineering areas are at the extreme negative

<sup>2</sup> J. B. Kruskal, How to use M-D-SCAL, a program to do multidimensional scaling and multidimensional unfolding, March 1968. This paper and the accompanying computer program can be obtained by writing to J. B. Kruskal, Bell Telephone Laboratory, Murray Hill, New Jersey 07974.

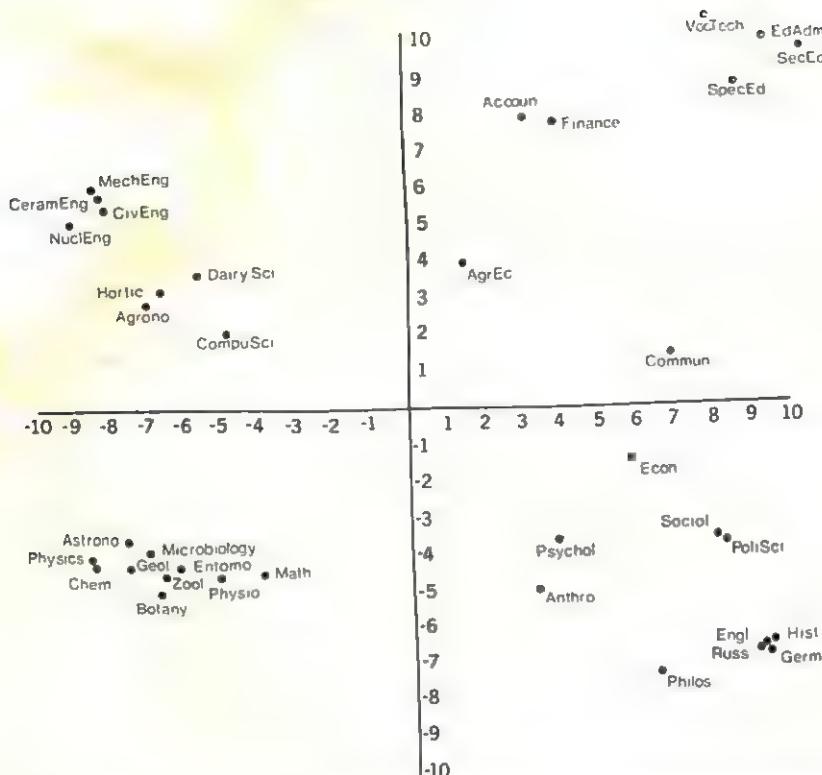


FIG. 1. Dimension I appears horizontally, and Dimension II appears vertically.

end, while humanities and education areas are at the extreme positive end. Biological areas are on the negative side, though closer to the origin than are the humanities. We thus have "hard" or science-oriented areas at one end of the dimension, social sciences toward the middle, and humanities at the other end of the dimension.

The second dimension (Figures 1 and 2) is a pure-applied dimension. At the extreme positive end are education areas. Accountancy, finance, and engineering areas are also at the positive end. On the negative end are physical sciences, mathematics, social sciences, languages, history, and philosophy. Unlike areas at the negative end of this dimension, those at the positive end are concerned with practical application of their subject matter.

The third dimension (Figures 2 and 3) appears to reflect the areas' concern with living or organic objects of study. Areas at the positive end all study such subject matter, while areas at the negative end do not. Thus, agricultural, biological, social science, and

education areas are high on the dimension. The first two of these groups involve study of all living systems, while the latter two groups are concerned primarily with the study of man. On the negative end of this dimension are all of the areas that do not study living things. These areas do not seem to be widely dispersed, and it appears that the only characteristics they have in common is the absence of biological objects of study.

#### *Scaling of Small College Data*

For the small college sample, solutions in six, five, four, and three dimensions were obtained, and each was rotated to principal axes to aid interpretation. Stress values for these solutions were .054, .087, .124, and .184 for the six- through three-dimensional solutions, respectively. The four-dimensional solution was chosen as the "best" solution because all four of its dimensions were interpretable, and its stress value was "good" (.124) according to Kruskal's suggested evaluations.

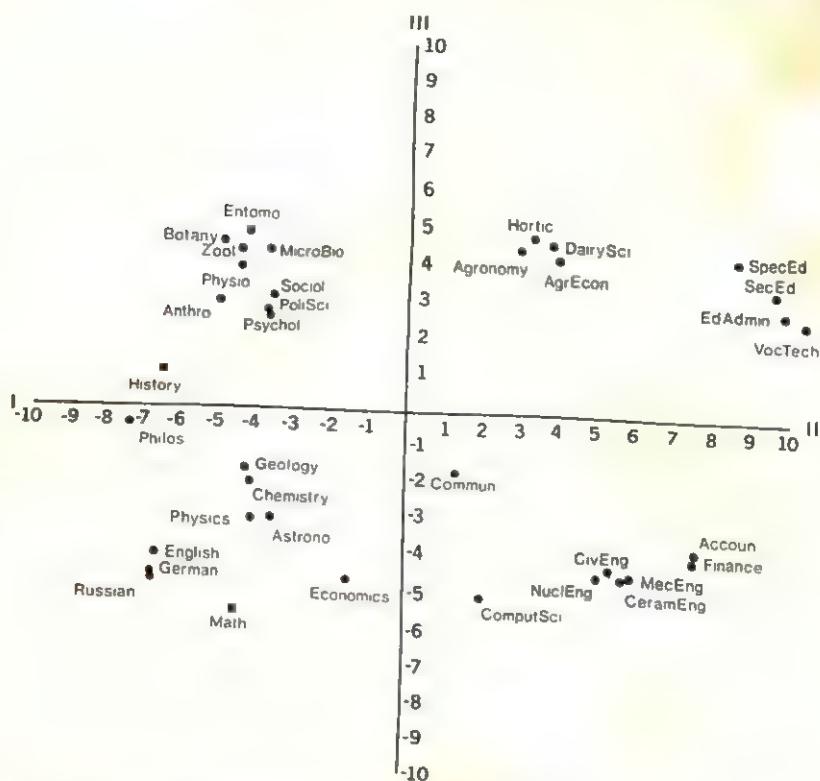


FIG. 2. Dimension II appears horizontally, and Dimension III appears vertically.

We may first ask if any of the dimensions of this solution are comparable to dimensions of the Illinois three-dimensional solution. Since 18 areas were common to both solutions, this question can be examined by correlating the positions of these areas on each dimension of the Illinois solution with their position on each dimension of the small college solution. Table 1 presents these correlations. The first dimension of the Illinois solution is virtually identical ( $r = .96$ ) to the first dimension of the small college solution. The dimension distinguishes hard sciences from social sciences and humanities. The second dimension of the Illinois solution is highly correlated ( $r = -.81$ ) with the third dimension of the small college solution. (The negative relationship is due to the inflection of the dimension on one solution and is of no consequence for interpreting the dimensions.) This dimension was interpreted in the Illinois solution as "concern with application." Visual inspection of the third dimension of the small college solution suggested the same interpretation. On the third

Illinois dimension, areas with biological or social objects of study are distinguished from other areas. This dimension is highly related to the fourth dimension of the small college solution ( $r = .89$ ). Thus, it appears that a dimension involving concern of areas with

TABLE 1  
CORRELATIONS BETWEEN THE THREE DIMENSIONS OF THE ILLINOIS SOLUTION AND THE FOUR DIMENSIONS OF THE SMALL COLLEGE SOLUTION FOR 18 AREAS COMMON TO BOTH SAMPLES

Small college dimension	Illinois dimension		
	I	II	III
(I)	.96	-.35	-.03
(II)	-.47	.16	-.36
(III)	-.13	-.81	-.20
(IV)	.09	.07	.89

# ANTHONY BIGLAN

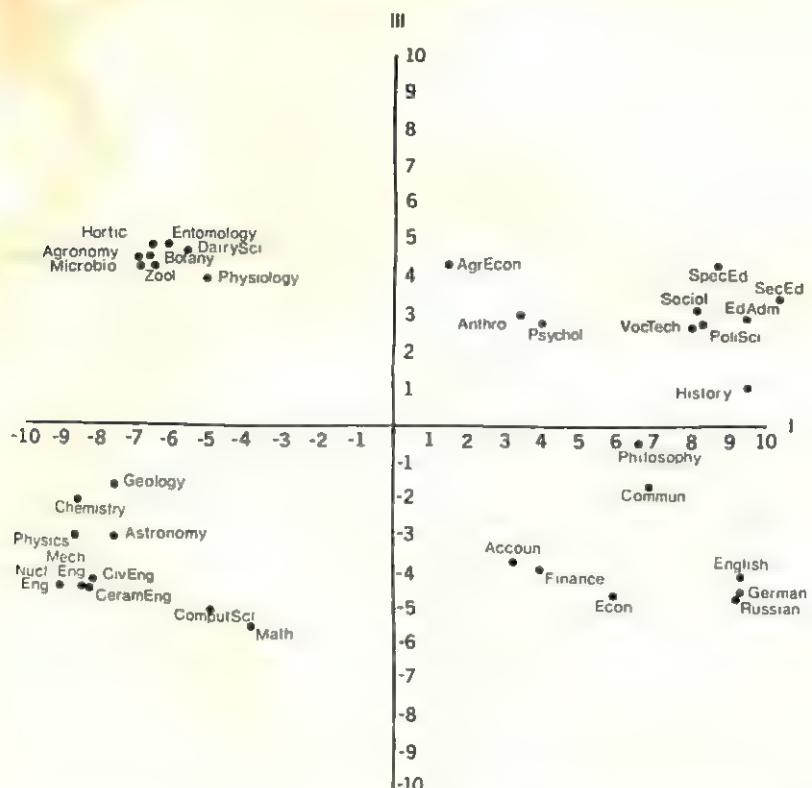


FIG. 3. Dimension I appears horizontally, and Dimension III appears vertically.

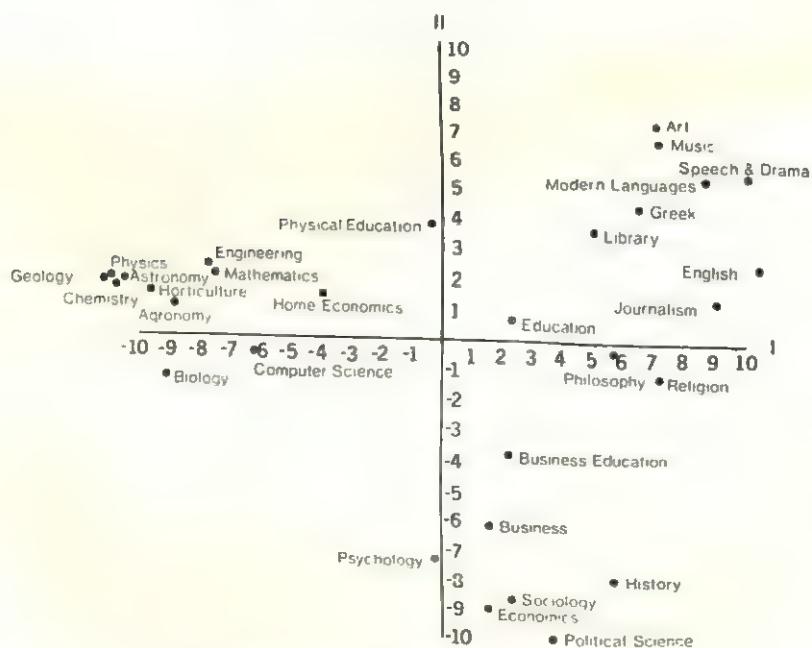


FIG. 4. Dimensions I and II of the small college solutions;  
Dimension I is the horizontal dimension.

biological or social processes is common to both solutions.

The second dimension of the small college solution is not strongly related to any of the Illinois dimensions. Figure 4 shows this dimension plotted against the first dimension of the small college solution. Art, music, speech and drama, and modern languages are at the positive end of this dimension, while social sciences such as political science, economics, and sociology are at the negative end. All of the areas that are a substantial distance from the origin are commonly found in liberal arts curriculae. Those at the positive end emphasize creative approaches to their subject matter, while those at the negative end emphasize empirical approaches. We may, therefore, tentatively label this dimension creative versus empirical liberal arts.

It is also useful to inquire about the overall similarity between the Illinois and small college solutions. This problem was examined by computing canonical correlations between the two solutions for the eighteen areas common to both. The three canonical correlations are .99, .92, and .88, indicating that the two solutions are highly similar.

#### *Attribute Analysis*

Interpretation of these dimensions becomes more clear when they are related to ratings of each area's attributes. Scholars at the small college rated each area on six bipolar adjectives. These ratings were averaged over all raters, and the average for each area was correlated with its position on each of the four dimensions obtained from the replication scaling. There were, thus, six attributes correlated with each of four dimensions. Table 2 presents these correlations.

Dimension I is correlated (.73) with the physical-nonphysical rating, indicating that the areas arrayed along this dimension differ in the extent to which they study physical objects. Two other attributes, biological-nonbiological and interesting-of no interest, were substantially related to the first dimension, but neither is so highly related to the dimension as to suggest a straightforward interpretation.

Dimension II is not strongly related to any of the attributes. It was suggested above that

TABLE 2  
CORRELATIONS BETWEEN DIMENSIONS OF ACADEMIC AREA SCALING (SMALL COLLEGE SAMPLE)  
AND ATTRIBUTE RATINGS ( $N = 30$ )

Attribute rating	Academic area dimension			
	I	II	III	IV
Pure-Applied	-.01	.04	-.82	-.09
Physical-Nonphysical	.73	-.26	.40	-.26
Biological-Nonbiological	-.52	-.03	-.15	.66
Interesting-				
Of no interest	.50	-.16	.26	.36
Traditional-				
Nontraditional.	-.22	-.15	-.51	-.00
Life science-				
Nonlife science	-.44	-.25	-.10	.68

this dimension involves creative versus empirical approaches to liberal arts. Dimension III was interpreted above as involving concern with application. This interpretation is supported by the correlation ( $r = -.82$ ) between this dimension and the pure-applied attribute.

Dimension IV distinguishes biological and social fields from other areas. The fourth column of Table 2 shows that both the biological-nonbiological and life science-nonlife science ratings are correlated with dimension IV. However, neither correlation is high enough to justify labeling the dimension according to either attribute. The problem is that neither attribute deals with the extent to which the area is concerned with social processes. Perhaps the best name for this dimension is "concern with life systems."

#### DISCUSSION

Three characteristics of academic subject matter are perceived by scholars in both a university and a small college setting. The most prominent dimension (in terms of the variance it accounts for) distinguishes hard sciences, engineering, and agriculture from social sciences, education, and humanities. A good shorthand label for the dimension is "hard-soft." The dimension appears to provide one kind of empirical support for Kuhn's (1962) analysis of the paradigm. By "paradigm" Kuhn refers to a body of theory which is subscribed to by all members of the field. The paradigm serves an important organizing

function; it provides a consistent account of most of the phenomena of interest in the area and, at the same time, serves to define those problems which require further research. Thus, fields that have a single paradigm will be characterized by greater consensus about content and method than will fields lacking a paradigm. Kuhn specifically designates physical and biological sciences as paradigmatic. He does not discuss agricultural and engineering areas, but they may also be considered to be paradigmatic, since they are grounded in their related pure fields. The areas at the extreme positive end—the humanities and education areas—are not paradigmatic. Rather, content and method in these areas tend to be idiosyncratic. The social sciences and business areas are also on the positive end of this dimension, but closer to the origin. These are fields that strive for a paradigm; but have yet to achieve one.

A second dimension underlying the way scholars view academic areas is the concern of the area with application to practical problems. Education, engineering, and agricultural areas are distinguished from hard sciences, social sciences, and humanities. The interpretation of this dimension is supported by its correlation with ratings of the areas on a pure-applied attribute dimension ( $r = -.82$ ,  $N = 30$ ). This dimension also appears to be used by scholars regardless of the kind of institution they are associated with.

Scholars also distinguish biological and social areas from those that deal with inanimate objects. This dimension also appears to be general to scholars in diverse institutions, since it was used by those at the University of Illinois and at a small liberal arts college. It is labeled "concern with life systems."

The one dimension that was not used by scholars at both institutions distinguished creative and empirical liberal arts areas. It is possible that this dimension did not appear in the Illinois solution because the areas that define the positive end of the dimension (art, music, and speech and drama) were not included in the Illinois judgment task. It is also possible that this dimension merely reflects the way that areas are grouped at the liberal arts college where we collected data.

This study has significance for at least two

aspects of the scientific investigation of scholarly endeavors. First, investigations of the role of social structure in scholarly work tend to be restricted to a single or a few academic areas (Gouldner, 1970; Menzel, 1962; Pelz & Andrews, 1966). The subject matter differences that have been described here show why it may be unwise to generalize such studies to other academic areas. A subsequent article (Biglan, 1973) is addressed to this problem. Relationships are examined between the subject matter characteristics identified in this study and the structure and output of university departments.

Second, the analysis is relevant to the study of the cognitive processes of different areas. Increasing emphasis is being given to the way in which both the content and methods of a field are linked to the cognitive and perceptual processes of its members. Kuhn (1962) has shown how changes in scientific theory can be understood as a process of cognitive reorganization on the part of people in the field. Consistent with this, Piaget (1971) draws parallels between the conceptual systems of science and basic aspects of cognitive development. The present analysis provides a systematic framework for exploring the role of cognitive processes in academic fields. Specifically, it suggests that the three most important dimensions for characterizing the "cognitive style" of an area concern its use of a paradigm, its attention to practical application, and its concern with life systems. Moreover, the analysis presented here suggests the degree to which styles are similar in different areas.

In summary, three dimensions appear to characterize the subject matter of academic areas in most institutions. The dimensions involve (a) the degree to which a paradigm exists, (b) the degree of concern with application, and (c) concern with life systems. These characteristics may have an important effect on the type of structure and output that a department has. Moreover, these dimensions may provide a useful framework for studying the cognitive style of scholars in different areas.

#### REFERENCES

- BIGLAN, A. Relationships between subject matter characteristics and the structure and output of

- university departments. *Journal of Applied Psychology*, 1973, 58, 204-213.
- CARROLL, J. D., & CHANG, J. J. Analysis of individual differences in multidimensional scaling via an  $N$ -way generalization of 'Eckart-Young' decomposition. Murray Hills, N. J.: Bell Telephone Laboratories, 1969. (Mimeo)
- GOULDNER, A. *The coming crisis in western sociology*. New York: Basic Books, 1970.
- KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 1964, 29, 1-27. (a)
- KRUSKAL, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, 29, 28-42. (b)
- KUHN, T. S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.
- MENZEL, H. Planned and unplanned scientific communication. In B. Barber & W. Hirsch (Eds.), *The sociology of science*. New York: The Free Press, 1962.
- MILLER, G. A. A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, 1969, 6, 169-191.
- ONCKEN, G. *Organizational control in university departments*. (Tech. Rep. No. 71-20) Seattle, Wash.: Organizational Research Group, University of Washington, June 1971.
- PELZ, D. C., & ANDREWS, F. M. *Scientists in organizations*. New York: Wiley, 1966.
- PIAGET, J. *Psychology and epistemology*. New York: Grossman, 1971.
- RAO, V. R., & KATZ, R. An empirical evaluation of alternative methods for the multidimensional scaling of large stimulus sets. Unpublished manuscript, Cornell University and University of Pennsylvania, 1970.
- RICHARDS, L. G. A multidimensional scaling analysis of judged similarity of complex forms from two task situations. *Perception and Psychophysics*, in press.
- SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, Parts I and II. *Psychometrika*, 1962 27, 125-140, 219-246.

(Received December 28, 1971)

# RELATIONSHIPS BETWEEN SUBJECT MATTER CHARACTERISTICS AND THE STRUCTURE AND OUTPUT OF UNIVERSITY DEPARTMENTS<sup>1</sup>

ANTHONY BIGLAN<sup>2</sup>

*University of Washington*

The social structure and output of scholars at the University of Illinois are examined in terms of the characteristics of their academic subject matter. On the basis of an earlier multidimensional analysis (Biglan, 1973) academic areas were clustered according to their (a) concern with a single paradigm (hard vs. soft), (b) concern with application (pure vs. applied), and (c) concern with life systems (life system vs. nonlife system). Depending on the characteristics of their area, scholars differed in (a) the degree to which they were socially connected to others, (b) their commitment to teaching, research, and service, (c) the number of journal articles, monographs, and technical reports that they published, and (d) the number of dissertations that they sponsored.

This article examines relationships between the characteristics of academic subject matter and the structure and output of university departments. Despite considerable attention to university organization in recent years, the possibility that the subject matter requires or contributes to particular kinds of organization has not been systematically evaluated. In an earlier article (Biglan, 1973), scholars' judgments identified three important features of academic subject matter. Academic areas differ according to (a) the existence of a single paradigm, (b) their concern with practical application, and (c) their concern with life systems. This study defines limits on the generality of organization studies that are restricted to a single academic area and calls attention to the dangers inherent in ignoring subject matter characteristics.

<sup>1</sup> This study was supported in part by the Office of the Executive Vice President and Provost, University of Illinois, and in part by the U.S. Office of Education, Department of Health, Education, and Welfare, Office of Education Bureau of Research 0-0340, Grant OEG-0 70,3447 (Fred E. Fiedler, principal investigator). The author would like to express special thanks to Gerald Oncken. In addition, thanks are expressed to Fred Fiedler, Lyle Lanier, Martin Ziegler, and David DeVries, along with more than 60 others, for their suggestions and support during this research.

<sup>2</sup> Requests for reprints should be sent to Anthony Biglan, Department of Psychiatry, University of Wisconsin, 427 Lorch Street, Madison, Wisconsin, 53706.

## UNIVERSITY DEPARTMENTS

### *Department Structure and Output*

*Social connectedness among faculty members.* Unlike departments in most formal organizations, university departments do not have clear lines of authority in which some members must answer to others. Oncken (1971) showed that the typical university department has a distribution of control that is egalitarian. In the absence of a clear, formal structure, informal relations among colleagues—their social connections—may be crucial to the department's functioning efficiently. Informal social connections also appear important for research activities, at least in the sciences. Hagstrom (1964) found teamwork to be characteristic of physical science research. In these areas, the scholar's informal relations with his colleagues are a prime source of technical information (Menzel, 1962) and appear to contribute to his scholarly productivity (Pelz & Andrews, 1966).

Despite the apparent importance of social connectedness among scholars, its extent in different academic areas has not been investigated. The present study examines whether social connectedness varies with the characteristics of academic subject matter. Of particular interest is the question of whether high social connectedness is characteristic of areas other than physical sciences. A second and equally significant question is whether social connectedness is positively associated

with scholarly productivity in areas other than the hard sciences. Despite the evidence just cited for such a positive relationship in hard science areas, the relationship between social connectedness and scholarly productivity has not been investigated in other areas.

Three aspects of scholars' social connectedness are examined in the present study. First, an individual may be connected to others in the sense that he likes working with them. Second, he may be connected by the extent to which others influence him. Finally, an individual is connected to others to the extent that he actually collaborates with them. Since teaching and research activities may engender different degrees of social connectedness, these three aspects of connectedness are examined separately for the two activities.

*Commitment to teaching, research, administration, and service.* Considerable controversy has raged in academia in recent years concerning the relative emphasis that should be placed on teaching and research. However, appropriate standards for these and other scholarly activities may depend on the nature of the area. What evidence exists indicates that the emphasis on, and significance of, teaching differs in physical and social science fields. Scholars in social sciences emphasize educating the whole student and evidence a more personal commitment to students than do those in physical sciences (Gamson, 1966; Vreeland & Bidwell, 1966). Although informative, these studies need to be extended and elaborated. First, we need to examine whether emphasis on research, administration, and service activities also differs according to academic area. Moreover, it is important to know if scholars in the various areas simply differ in preferences for these activities or if they actually spend different amounts of time on them. Both of these questions are examined in the present study. The commitment of scholars in different areas to teaching, research, administration, and service are examined in terms of (a) liking for the activity and (b) the amount of time they actually spend on the activity.

*Scholarly output.* The evidence is rather strong that different measures of scholarly output do not converge (Smith & Fiedler, 1971). Thus, a variety of output measures are

included in the present study. In the case of research, the quantity of monographs, journal articles, and technical reports are included as well as a measure of journal article quality that is based on the rated quality of the journal in which it is published. The effectiveness of graduate training at the doctoral level is indexed by ratings of the quality of the first jobs that graduate students obtain upon completing their degrees and the number of doctoral dissertations sponsored. Unfortunately, no index of undergraduate teaching effectiveness was available.

Despite research on relationships among scholarly output measures (cf. Cole & Cole, 1967), the question of whether these measures differ systematically with academic area appears not to have been examined. The answer to this question has important implications for the way we shall evaluate faculty members. If, for example, faculty members produce different numbers of monographs depending on their area, then we may want to weight monographs differently when evaluating scholars in different areas.

## METHOD

Data on department structure and output were collected at the Urbana campus of the University of Illinois in the spring of 1968. The university is a large, state-supported institution with an extensive commitment to research and graduate education. Most academic disciplines are represented on the Urbana campus; there are over 100 distinct curricula.

In the early stages of research, data were collected on the organization of 47 departments. Since one purpose of our research was the study of the characteristics of successful graduate programs, only departments granting PhDs were included in the sample.

### Sources of Data

The chief sources of structure and output data were questionnaires, archival records, and faculty members' judgments of certain outputs. The questionnaire asked scholars about the structure of their social relations and their commitment to teaching, research, administration, and service. Department heads in 47 departments were contacted through the Dean of the Graduate College. They were asked to fill out the questionnaire and to ask the members of their department to do the same. The remaining members of the faculty received their questionnaires by mail. Response rates within the departments ranged from 19% to 100%, and the overall rate was 55%. Because of their low response rate, some departments were deleted from the present

TABLE 1  
OPERATIONAL MEASUREMENT OF SOCIAL CONNECTEDNESS AND COMMITMENT VARIABLES

Variable	Description
Social connectedness	
Number of others—like to work with	Respondents to the questionnaire listed people they said they liked to work with on teaching, research, and administration. The number of people named for each of these tasks was the measure.
Number of sources of influence	Respondents were asked to indicate the individuals and groups who influenced their research goals and teaching procedures. The number of sources indicated was the measure.
Collaboration	Respondents to the questionnaires indicated the number of fellow faculty members with whom they worked directly on research and teaching.
	A second measure of research collaboration was obtained by tabulating the number of coauthorships each faculty member had on his journal articles.
Commitment Preferences	Questionnaire respondents were asked to distribute 100 points among the following tasks in accordance with their preferences for each task: teaching, research, department administration, university administration, and service.
Time allocation	In a similar manner, respondents distributed 100 points among these tasks to indicate the proportion of time they spent on each. Since respondents also indicated the number of hours they spent on all university work, it was possible to devise measures of time spent on each activity.

study. The average response rate of departments retained in this study was 65%.<sup>3</sup>

Archival records provided data about publication quantity and the first jobs which finishing graduate students obtained. An official university pamphlet entitled *Publications of the Faculty* is published annually. It lists all monographs, journal articles, technical reports, and dissertations sponsored by faculty members during the preceding year. Departmental records provided information on the specific jobs obtained by all

graduate students who had completed their PhDs in the years 1964-1968.

It was important to obtain measures of the quality of jobs and publications as well as their quantity. Our approach to this problem was to ask faculty members to rate the quality of graduate students' first jobs and the journals in which the scholars in our sample had published.

#### *Operational Measurement of Variables*

Table 1 lists social connectedness and commitment variables and describes the specific operations involved in deriving each. All but one of these variables was derived from the questionnaire.

Measures of publication quantity were tabulated for each faculty member who received a questionnaire. The quantity of four kinds of publications was tabulated: monographs, journal articles, dissertations sponsored by the scholar, and technical reports.

A separate paper (Biglan, Oncken, & Fiedler, 1971) presents a detailed description of the development of the journal article and first-job quality measures and presents evidence relevant to their reliability and validity. Briefly, the measure of journal article quality was derived for each questionnaire respondent who had published at least one article in the period 1964-1967. The measure was based on the ratings of journal quality that were described above. Each of the journals in which the scholar had published during the 4-year period of interest was noted, and the quality score for that journal was recorded. Then the quality scores

<sup>3</sup> Comparison of respondents and nonrespondents for all departments included in the original sample indicated that nonrespondents were more peripheral members of the department. They were less likely to have advanced degrees, had a smaller percentage of their time devoted to the university and department, and spent more of their time on teaching. Although these differences were statistically significant, none accounted for more than 2.1% of the variance. In another analysis, the relationships between subject matter characteristics and return rate were examined by correlating the return rate of each department with three measures of the characteristics of the department's subject matter (Biglan, 1973). Response rate was significantly related to the degree of concern with application ( $r = .50$ ,  $p < .05$ ), indicating that departments in applied areas had higher rates of response. The two other measures—existence of a paradigm and concern with life systems—were not significantly related to return rate ( $r = -.13$  and  $r = .03$ , respectively).

TABLE 2  
CLUSTERING OF ACADEMIC TASK AREAS IN THREE DIMENSIONS

Task area	Hard		Soft	
	Nonlife system	Life system	Nonlife system	Life system
Pure	Astronomy Chemistry Geology Math Physics	Botany Entomology Microbiology Physiology Zoology	English German History Philosophy Russian Communications	Anthropology Political science Psychology Sociology
Applied	Ceramic engineering Civil engineering Computer science Mechanical engineering	Agronomy Dairy science Horticulture Agricultural economics	Accounting Finance Economics	Educational administration and supervision Secondary and continuing education Special education Vocational and technical education

were summed and divided by the number of journal articles the scholar had published. An index of the quality of the first jobs of each scholar's graduate students was developed in essentially the same manner. A score for the quality of each job was obtained by averaging the judges' ratings. The final job quality measure for the scholars was then derived by averaging these quality scores for all of the jobs that the particular scholar's graduate students had obtained.

### Analysis of Data

In an earlier article (Biglan, 1973), a multidimensional analysis of 36 academic subject areas was presented. Three dimensions were derived from the judgments of scholars at the University of Illinois. The dimensions involved (a) the existence of a single paradigm (hard-soft), (b) concern with practical application (pure-applied), and (c) concern with life systems. It is possible to cluster areas on the basis of their position on each of these three dimensions. Table 2 presents an organization of areas in eight clusters. The table lists the areas included in each cluster. Each cluster centroid is located in a different octant of the three-dimensional space and can thus be characterized according to whether it is hard or soft, pure or applied, and concerned with life systems or not.

This clustering suggests an analysis of variance approach to our examination of relationships between area characteristics and department structure and output. Specifically, a three-way analysis of variance design corresponding to Hard versus Soft  $\times$  Pure versus Applied  $\times$  Life System versus Nonlife System was employed in the analysis of structure and output data. Thus, each subject's data falls into one of the octants of this three-way design. In examining the way in which area characteristics mediate relationships between social connectedness and scholarly output, a four-way analysis of variance was performed. Here the

four factors correspond to the high versus low social connectedness by the three area factors just mentioned.

## RESULTS

### Hard versus Soft Areas

*Social connectedness.* Hard and soft areas differ significantly on one of three measures of social connectedness in teaching and on three of the four measures of social connectedness in research. In each case, it is the hard areas that are higher in connectedness. For teaching activities, scholars in hard areas report greater collaboration with fellow faculty members ( $\bar{X}_H = .66$ ) than do those in soft areas ( $\bar{X}_S = .29$ ,  $F = 17.52$ ,  $df = 1/429$ ,  $p < .01$ ). There were no differences in the number of people with whom they reported liking to work on teaching or in the number of reported sources of influence on the courses they teach. For research activities, scholars in hard areas like to work with significantly more people on research ( $\bar{X}_H = 1.93$ ) than do those in soft areas ( $\bar{X}_S = 1.36$ ,  $F = 14.29$ ,  $df = 1/584$ ,  $p < .01$ ). Similarly, hard area scholars report more sources of influence on their research goals ( $\bar{X}_H = 2.12$ ,  $\bar{X}_S = 1.70$ ,  $F = 21.74$ ,  $df = 1/569$ ,  $p < .01$ ). The extent to which scholars collaborate with other faculty members on research did not differ according to the hard-soft distinction or according to any of the other area characteristics. Many

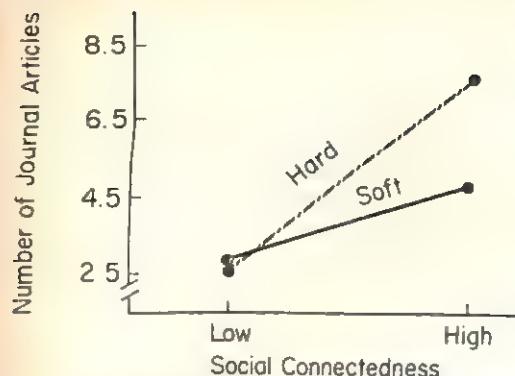


FIG. 1. Interaction between social connectedness and the hard-soft factor on journal article publications.

respondents appeared not to understand the instructions to this question. As a result, a second measure of research collaboration, the number of journal coauthors, was included in the study. Analysis of this measure showed that hard area scholars have a significantly greater number of coauthors ( $\bar{X}_H = 5.67$ ) than do their soft area ( $\bar{X}_S = .63$ ) counterparts ( $F = 47.48, df = 1/473, p < .01$ ).

**Commitment.** Hard and soft area scholars differ significantly in their commitment to teaching and research. As compared with hard areas, scholars in soft areas indicate a greater preference for teaching ( $\bar{X}_H = 37.1, \bar{X}_S = 48.7, F = 41.63, df = 1/620, p < .01$ ) and actually spend more time on it ( $\bar{X}_H = 19.1, \bar{X}_S = 26.4, F = 42.29, df = 1/603, p < .01$ ). For research, the situation is just the reverse. Hard area scholars show significantly greater preference for research than do those in soft areas ( $\bar{X}_H = 41.1, \bar{X}_S = 31.8, F = 22.89, df = 1/620, p < .01$ ) and actually spend more time on it ( $\bar{X}_H = 23.0, \bar{X}_S = 15.1, F = 37.97, df = 1/603, p < .01$ ). The analyses also revealed three-way interactions among the three area characteristics (i.e., hard-soft, pure-applied, life system-nonlife system) in both preference for ( $F = 21.08, df = 1/620, p < .01$ ) and time spent on research ( $F = 13.79, df = 1/603, p < .01$ ). These interactions indicate that differences between hard and soft areas in preference for, and time spent on research, are greatest in applied life system areas (agriculture and education) and pure nonlife system areas (physical sciences and humanities). In other words, the greatest differences on these variables are between

agriculture and education and between physical sciences and humanities.

**Scholarly output.** The rate of publication of monographs and journal articles are both related to the hard-soft distinction. Scholars in hard areas produce significantly fewer monographs than do those in soft areas ( $\bar{X}_H = .08, \bar{X}_S = .28, F = 14.54, df = 1/473, p < .01$ ), and they produce significantly more journal articles ( $\bar{X}_H = 6.21, \bar{X}_S = 2.72, F = 25.31, df = 1/473, p < .01$ ) than soft area scholars. Caution, however, is appropriate in considering this last result. Since, as was shown above, the incidence of joint authorship is greater in hard areas and since journal articles were credited to the scholar when he was not first author, the greater incidence of journal articles in hard areas must be in part due to the same article being credited to more than one scholar.

**The relationship between social connectedness and scholarly output.** A significant interaction was found between social connectedness and the hard-soft factor in their effects on journal article publication ( $F = 6.22, df = 1/473, p < .01$ ). This interaction is shown in Figure 1. It indicates that social connectedness is more strongly related to journal article publication in hard areas than it is in soft areas. A second interaction between social connectedness and the hard-soft factor indicates that social connectedness and scholars' technical report publication are positively related in hard areas but negatively related in soft areas ( $F = 4.32, df = 1/473, p < .01$ ).

Two other significant interactions are appropriately presented here. The social con-

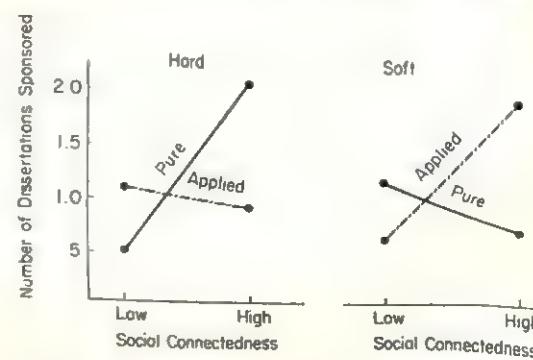


FIG. 2. Interaction among social connectedness, the hard-soft factor, and the pure-applied factor on number of dissertations sponsored.

nectedness, hard-soft, and pure-applied factors significantly interacted in their relationship to the number of dissertations that the scholars in our sample sponsored ( $F = 13.91$ ,  $df = 1/473$ ,  $p < .01$ ). Figure 2 illustrates this interaction. Positive relationships between connectedness and dissertations sponsored occurred in hard, pure areas such as physics and physiology and in soft, applied areas such as education and finance. An almost identical interaction occurred for the quality of graduate students' first jobs ( $F = 7.17$ ,  $df = 1/473$ ,  $p < .01$ ). Job quality is positively related to social connectedness in hard, pure areas and in soft, applied areas; job quality and connectedness are unrelated in the remaining areas.

### Pure versus Applied Areas

*Social connectedness.* Pure and applied areas differ significantly on one of three measures of teaching connectedness and two of four measures of research connectedness. Scholars in applied areas like to work with significantly more people on teaching than do scholars in pure areas ( $\bar{X}_A = 1.30$ ,  $\bar{X}_P = .93$ ,  $F = 10.13$ ,  $df = 1/584$ ,  $p < .01$ ). Similarly, applied area scholars like to work with more people on research than do those in pure areas ( $\bar{X}_A = 1.88$ ,  $\bar{X}_P = 1.41$ ,  $F = 9.98$ ,  $df = 1/584$ ,  $p < .01$ ). And they report more sources of influence on their research goals than do the pure area scholars ( $\bar{X}_A = 2.18$ ,  $\bar{X}_P = 1.63$ ,  $F = 37.47$ ,  $df = 1/569$ ,  $p < .01$ ). A significant interaction between the pure-applied and hard-soft factors was also found for number of sources of influence on research goals ( $F = 14.44$ ,  $df = 1/569$ ,  $p < .01$ ). It showed that the difference between pure and applied areas on this variable is larger for hard areas (e.g., physics vs. engineering) than it is for soft areas (e.g., education vs. English).

*Commitment.* Scholars in pure areas like research activities more than do those in applied areas ( $\bar{X}_A = 33.3$ ,  $\bar{X}_P = 39.7$ ,  $F = 11.02$ ,  $df = 1/620$ ,  $p < .01$ ). However, according to our results for time spent, pure area faculty do not actually spend more time on research. Applied area scholars like service activities more than do those in pure areas ( $\bar{X}_A = 7.8$ ,  $\bar{X}_P = 3.4$ ,  $F = 33.81$ ,  $df = 1/603$ ,  $p < .01$ ) and actually spend more time on

them ( $\bar{X}_A = 4.4$ ,  $\bar{X}_P = .26$ ,  $F = 12.75$ ,  $df = 1/603$ ,  $p < .01$ ). A significant three-way interaction on preference for service shows that the main effect difference between pure and applied scholars' preference is primarily due to the high degree of liking for service that was reported by individuals in education (soft, applied, life system fields) and engineering (hard, applied, nonlife system fields) areas ( $F = 15.49$ ,  $df = 1/620$ ,  $p < .01$ ). A similar result occurred for the amount of time actually devoted to service, but it was only significant at the .05 level.

*Scholarly output.* Pure and applied areas differ in the production of technical reports and the rated quality of their graduate students' first jobs. Applied area scholars publish more technical reports ( $\bar{X}_A = .46$ ,  $\bar{X}_P = .16$ ,  $F = 6.64$ ,  $df = 1/473$ ,  $p < .01$ ), and the rated quality of graduate students' first jobs is higher in applied areas than it is in pure areas ( $\bar{X}_A = 5.82$ ,  $\bar{X}_P = 4.85$ ,  $F = 10.30$ ,  $df = 1/75$ ,  $p < .01$ ).

*The relationship between social connectedness and scholarly output.* The relationship between social connectedness and rate of monograph publication differs, depending on whether the area is pure or applied ( $F = 4.09$ ,  $df = 1/473$ ,  $p < .01$ ). In pure areas, connectedness is positively related to monograph publication, while in applied areas the scholars' social connectedness makes no difference. An interaction was found among the social connectedness, pure-applied, and life system factors in their relationship to the technical report publication of scholars. Social connectedness and technical report output are positively related in applied life system fields (education, agriculture), negatively related in pure life system areas (life and social sciences), and unrelated in other areas ( $F = 4.25$ ,  $df = 1/473$ ,  $p < .01$ ).

### Life System versus Nonlife System Areas

*Social connectedness.* Scholars in life system and nonlife system areas differ in the number of people with whom they like to work on teaching. Those in life system areas like to work with significantly more people ( $\bar{X}_{LS} = 1.28$ ,  $\bar{X}_{NL} = .94$ ,  $F = 8.85$ ,  $df = 1/584$ ,  $p < .01$ ). Moreover, there is a significant three-way interaction for the effects of area

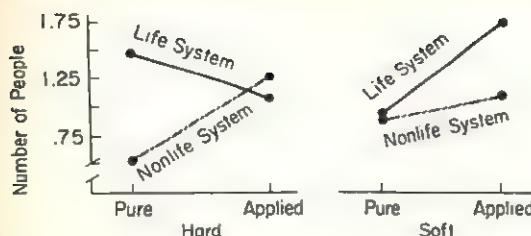


FIG. 3. Three-way interaction for number of people with whom respondent likes to work on teaching.

characteristics on the number of people with whom scholars like to work on teaching activities ( $F = 12.43$ ,  $df = 1/584$ ,  $p < .01$ ). The interaction is illustrated in Figure 3. It appears due to the differences between life system and nonlife system areas in hard, pure areas and in soft, applied areas. In both sets of areas, scholars in life system areas (i.e., life sciences and education) report liking to work with more people on teaching than do their counterparts in nonlife system areas (physical sciences and humanities).

The life system factor is related to only one of the four measures of research connectedness. Scholars in life system areas report significantly more sources of influence on their research goals than do scholars in nonlife system areas ( $\bar{X}_{LS} = 2.03$ ,  $\bar{X}_{NL} = 1.79$ ,  $F = 6.94$ ,  $df = 1/569$ ,  $p < .01$ ).

**Commitment.** Life system and nonlife system areas differ significantly on both measures of commitment to teaching, but they do not differ in commitment to any other scholarly activities. Scholars in life system areas indicate that they like teaching less than do scholars in nonlife areas ( $\bar{X}_{LS} = 38.7$ ,  $\bar{X}_{NL} = 47.6$ ,  $F = 26.40$ ,  $df = 1/620$ ,  $p < .01$ ). And, the life system scholars actually spend less time on teaching ( $\bar{X}_{LS} = 20.2$ ,  $\bar{X}_{NL} = 26.3$ ,  $F = 21.50$ ,  $df = 1/603$ ,  $p < .01$ ) than their nonlife counterparts. A significant interaction ( $F = 9.96$ ,  $df = 1/603$ ,  $p < .01$ ) among all three area factors showed that time spent on teaching is particularly small in agricultural areas (hard, applied life system areas).

**Scholarly output.** Life system areas did not differ significantly from nonlife system areas on any of our measures of scholarly output.

**Relationships between social connectedness and scholarly output.** Significant interactions occurred between social connectedness and the

life system factor as they are related to the number of dissertations sponsored ( $F = 6.91$ ,  $df = 1/473$ ,  $p < .01$ ) and the quality of graduate students' first jobs ( $F = 8.57$ ,  $df = 1/473$ ,  $p < .01$ ). Social connectedness is positively related to both of these output measures in areas that do not involve life systems, but is not related to them in life system areas.

## DISCUSSION

### *The Existence of a Paradigm*

The term "paradigm" refers to a body of theory that is subscribed to by all members of a field (Kuhn, 1962). The paradigm serves important organizing functions; it provides a consistent account of most of the phenomena of interest in the area and, at the same time, defines problems which require further study. Fields that have a single paradigm are characterized by greater consensus about appropriate content and method than are nonparadigmatic fields.

The present study suggests that a paradigm also permits structural and output features to develop that are not possible in nonparadigmatic areas. The paradigm permits greater social connectedness among scholars, particularly on their research. The common framework of content and method which it provides for the members of the field means that their attempts to work together will not be hindered by differences in orientation. In nonparadigmatic fields, on the other hand, scholars must work out a common definition of problems and method of approach before they can begin to work together. Our findings concerning social connectedness are that output relationships suggest that the paradigm may even *require* social connectedness in a way not true of soft or nonparadigmatic areas. Social connectedness is related more positively to both journal article and technical report publication in hard areas than it is in soft areas. Menzel's (1962) studies of physical sciences suggest that colleagues of the hard area scholar enhance his productivity by providing him with important technical information relevant to work on the paradigm. Connectedness may also be more highly related to scholarly output in paradigmatic areas because the paradigm permits research problems to be efficiently

broken into subproblems with confidence that the results for each part can be reintegrated.

The paradigm also appears to permit a more abbreviated form of scholarly communication. Compared to scholars in soft or non-paradigmatic areas, those in hard or paradigmatic areas publish fewer monographs and more journal articles. In paradigmatic areas, it is not necessary to provide detailed descriptions of the content and method that underlie a piece of research; these are understood by anyone familiar with the paradigm. In this case, journal articles, with their restrictions on length, provide an appropriate means of communication. In the soft areas, where paradigms are not characteristic, the scholar must describe and justify the assumptions on which his work is based, delimit his method or approach to the problem, and establish criteria for evaluating his own response to the problem. Such an undertaking requires a monograph-length work.

The paradigm may also account for the differences between hard and soft areas in commitment to teaching and research. The greater commitment of hard area scholars to research may be because important graduate training takes place in the research setting. As Kuhn (1962) suggests, budding scholars must be socialized to the regnant paradigm. One way for this to occur is for the graduate student in a hard area to work with a faculty member on his research. In nonparadigmatic areas, research is more independent and idiosyncratic (cf. the smaller social connectedness on research in soft areas). Thus, the faculty member will have less need for graduate research assistants, and at the same time, the graduate student will probably profit more from independent study than he will from working under a faculty member.

#### *Concern with Application*

Concern with application apparently requires a number of things of the individuals in a department. These include commitment to service activities, publication of technical reports, and a generally more socially connected collegial structure. The applied area scholar indicates a greater liking for service activities and actually spends more time on them. Perhaps as a compensation for this

commitment, scholars in applied areas report less liking for research activities than do their colleagues in pure areas. The service function of applied areas is also evident in the finding that scholars in applied areas publish more technical reports than their pure area colleagues. Presumably, technical reports provide an ideal format for communicating detailed research results to the groups and individuals who are serviced by applied areas.

Emphasis on the practical value of the scholar's work apparently leads him to rely on the evaluation of others. Compared with scholars in pure areas, those in applied areas report liking to work with more people on both research and teaching activities. And, applied area scholars report that their research goals are influenced by more sources. Examination of questionnaire responses indicated that many of these sources are outside agencies. This is particularly true in agricultural and engineering areas.

At least for some applied areas, it appears that the scholar's social connections to outside agencies increase the likelihood of his producing technical reports. Thus, in applied areas such as education and agriculture, social connectedness is related positively to the rate of technical report publication. In such pure areas as life and social sciences, these variables are related negatively, and in all remaining areas they are unrelated. One reason for these findings could be that when scholars in education and agriculture areas are high on our social connectedness measure, it is because they are connected to outside agencies which also encourage the scholars to write technical reports. In the social and life sciences, however, the scholar who scores high in social connectedness is probably connected to his colleagues. Such contacts would detract from, rather than enhance, his or her production of consumer-oriented technical reports.

#### *Concern with Life Systems*

The most distinctive characteristics of life system areas involve their graduate training. In many life system areas, this function appears to be performed by faculty members acting as a committee of the whole. Scholars in these areas report liking to work with more people on teaching activities. In nonlife areas,

the social connectedness of scholars is related positively to the number of dissertations they sponsor and the quality of their graduate students' first jobs. This is most likely because the scholars' connections help him find good jobs for students and this enhances his attractiveness as a sponsor of dissertations. However, in life system areas, social connectedness is not related to the sponsoring of dissertations or to first-job quality. Anecdotal evidence indicates that in many of these departments, the graduate student's work is periodically reviewed by a committee of faculty members. Moreover, job placement tends to be conducted by the central administration of the department. These factors would tend to diminish the importance of the social connections of the student's dissertation adviser.

In addition to these features, life system areas evidence less commitment to teaching activities. They like them less and spend less time on them than scholars in nonlife areas do. It may be that, like hard areas, life system areas train their graduate students in research settings. This is known to be the case for most life sciences at Illinois.

One characteristic of life system areas that does not involve graduate training is the influence on scholars' research goals. Individuals in life system areas are influenced by more people than are those in nonlife system areas. Examination of the questionnaires indicated that this is primarily a matter of the influence of outside agencies. It is possible that society has a more immediate and pressing concern for the products of research in these fields; fields such as education and life sciences are more directly relevant to the needs of large numbers of people. Hence, agencies outside the university attempt to shape directly the research being done in these fields.

### *Some Implications*

The findings of this study have important implications for the conduct of research on universities and for our procedures and practices in evaluating university faculty members.

*Research on universities.* The present study suggests the inadvisability of at least two approaches to studying university organizations. One approach is to collect organizational

data in a variety of fields and ignore area differences (Hill & French, 1967) in analyzing relationships among variables. This procedure is likely to mask different relationships in different areas. For example, for the data of the present study collapsed over area, we find a slight positive relationship between social connectedness and (a) rate of journal article publication ( $F = 3.99$ ,  $df = 1/473$ ,  $p < .01$ ) and (b) number of dissertations sponsored ( $F = 4.77$ ,  $df = 1/473$ ,  $p < .01$ ). But, as the results presented earlier show, the relationship between connectedness and these output variables may be significantly different, depending on the area. Thus, lumping together data from different areas may provide an inaccurate account of the organization of specific areas.

A second approach to organizational studies in universities is to restrict them to one or a few academic areas. This isn't bad in itself, but the findings presented here suggest that such studies will not be generalizable to dissimilar academic areas. For example, studies of collegial relations in the physical sciences indicate that social connectedness is high in these fields (Hagstrom, 1964) and that it enhances scholarly productivity (Menzel, 1962; Pelz & Andrews, 1966). The present study places distinct limits on the generality of these findings; it suggests that they hold also for engineering, agricultural, and life science areas, but not for such soft areas as education, humanities, and social sciences.

*Evaluation of faculty members.* The results of this study show that universitywide standards for the evaluation of faculty members will not be possible. To begin with, areas differ in their norms concerning commitment to teaching, research, and service. Hard areas evidence a greater commitment to research and a lesser commitment to teaching when compared with soft areas. Similarly, service is a distinctly more significant activity in applied areas than it is in pure areas. Thus, when we establish standards for evaluating the scholar's work, we shall first need to consider the relative importance of each of these scholarly activities in his or her area. Similar considerations arise when we examine the ways in which scholarly output is related to academic area. Hard area scholars publish more journal

articles and fewer monographs than do those in soft areas; applied area scholars publish more technical reports than do pure area scholars. In light of these findings it would be a mistake to give a journal article, monograph, or technical report the same weight when evaluating scholars in different areas. In sum, it appears that any attempt at universal standards for academia will impose a uniformity of activity and output which is inconsistent with the particular subject matter requirements of specific areas.

### SUMMARY

The structure and output of university departments are related to three characteristics of academic subject matter. The existence of an agreed upon paradigm in an area provides a structured framework that appears to encourage certain forms of organization. Compared to nonparadigmatic areas, those with a paradigm evidence greater social connectedness on research activities, greater commitment to research, less commitment to teaching, the publication of more journal articles, and the publication of fewer monographs. Moreover, social connectedness is positively related to journal article and technical report publication in paradigmatic areas, but this is not true of other areas. The organization of applied areas is distinct from that in pure areas. Applied areas evidence a greater commitment to service activities, a higher rate of technical report publication, and a greater reliance on colleague's evaluations. In contrast to nonlife system areas, scholars in life system areas appear to function as a group in training their graduate students and evidence a generally smaller commitment to teaching activities. Moreover, the public's interest in life system research is suggested by the greater influence of outside agencies on the research goals of life system scholars.

These results point to the need to consider subject matter characteristics in studying

academic organizations. They define limits on the extent to which studies in one area can be generalized to areas whose subject matter is different and indicate why studies of academic organizations should not lump together data that come from different areas. Finally, the study points to the need for evaluative standards that are appropriate to the particular activities and outputs of the academic area.

### REFERENCES

- BIGLAN, A. The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, 1973, 58, 195-203.
- BIGLAN, A., ONCKEN, G. R., & FIEDLER, F. E. *Convergency among academic outputs as a function of academic area*. (Tech. Rep. No. 71-26) Seattle, Wash: University of Washington, Organizational Research Group, 1971.
- COLE, S., & COLE, J. R. Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, 1967, 32, 377-390.
- GAMSON, Z. Utilitarian and normative orientations toward education. *Sociology of Education*, 1966, 39, 46-73.
- HAGSTROM, W. O. Traditional and modern forms of scientific teamwork. *Administrative Science Quarterly*, 1964, 9, 241-263.
- HILL, W. W., & FRENCH, W. L. Perception of power of department chairmen by professors. *Administrative Science Quarterly*, 1967, 11, 548-574.
- KUHN, T. S. *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1962.
- MENZEL, H. Planned and unplanned scientific communication. In B. Barber & W. Hirsch (Eds.), *The sociology of science*. New York: The Free Press, 1962.
- ONCKEN, G. *Organizational control in university departments*. (Tech. Rep. No. 71-20) Seattle, Wash.: University of Washington, Organizational Research Group, 1971.
- PELZ, D. C., & ANDREWS, F. M. *Scientists in organizations*. New York: Wiley, 1966.
- SMITH, R., & FIEDLER, F. E. Measurement of scholarly work in academic institutions: A review of the literature. *Educational Record*, 1971, 52, 225-232.
- VREELAND, R. S., & BIDWELL, C. E. Classifying university departments: An approach to the analysis of their effects upon undergraduates' values and attitudes. *Sociology of Education*, 1966, 38, 237-254.

(Received December 28, 1971)

## STATISTICAL ACCURACY AND PRACTICAL UTILITY IN THE USE OF MODERATOR VARIABLES

CRAIG C. PINDER<sup>1</sup>

*University of Minnesota<sup>2</sup>*

Using Ghiselli's (1956) prediction of predictability technique, two moderator variables—one empirically identified and the other, a hypothetical and perfectly accurate variable—were used in conjunction with a test battery to predict the criterion scores in a sample of customer engineers. The trade-off relationships between sample sizes, predictive accuracy, practical utility, and selection costs were explored, and many principles relating to moderated selection strategies were demonstrated. Improvements in multiple  $R$  and  $\sigma_{est}$  were gained by using the moderators. However, analysis of the average performance scores of subjects selected with the moderators suggested that the loss of sample size precluded the usual benefits derived through the use of small selection ratios.

Recent interest in the problem of improving the predictive accuracy of psychological tests in selection strategies through the use of "moderator" or "modifier" techniques has generated a myriad of proposed methods and an almost equal number of discrediting or qualifying criticisms and counterproposals. Zedeck (1971) has provided a comprehensive summary and integration of this literature.

One of the characteristics common to many strategies involving moderator variables, and among the most frequently criticized characteristics of these techniques, is the problem of the loss of utility that is encountered as the result of the discarding of "unpredictable" individuals. The criticism raised by McNemar (1969), for example, of the quadrant-analysis method proposed by Hobert (1965) and by Hobert and Dunnette (1967) is a case in point. Although predictive accuracy (in terms of improved multiple  $R$  and/or decreased standard error of estimate) may accrue from the discarding of unpredictable recruits and using a given test battery only to select personnel from a special subgroup of the total sample, the real benefit derived from the use of the battery in conjunction with such a controlled selection strategy will decrease as the proportion of the

sample discarded increases, and the costs of finding alternative strategies for the discards increases. As McNemar (1969) noted, a 100% hit rate is possible if we refuse to make statistical predictions for 93% of our cases, given a test with a validity coefficient of .70. The problem remains, what do we do with the "unpredictables?" In those cases where our selection ratios are very low there may be no problem, but in most practical situations this is not the case. Moreover, insofar as our means of determining who to test and who to discard before testing are not perfectly accurate, there will be an additional element of error and cost in the form of false negatives before testing, if such individuals are not hired into the organization by alternate selection procedures.

The problems raised here are not peculiar to the off-quadrant approach, but characterize any strategy that involves discarding subgroups of recruits before selection. Ghiselli's (1956) method of prediction of predictability is another technique that involves the same problems and costs.

Although many proponents, critics, and reviewers of the moderator literature have repeatedly raised and discussed these problems, there has been an obvious lack of empirical data demonstrating the trade-off relationships between improved accuracy and decreased net utility encountered in using these moderated techniques. The present study is an attempt to demonstrate the relationships between sample size, validity, predictive accuracy, and utility in selecting proficient employees through

<sup>1</sup> The author is grateful to George T. Milkovich and George W. England of the University of Minnesota Industrial Relations Center for their comments and suggestions during the study, and to Walter W. Tornow of Control Data Corporation for the use of his data.

Requests for reprints should be sent to Craig C. Pinder, 1063 Warren Road, Ithaca, New York 14850.

<sup>2</sup> Presently at the New York State School of Industrial and Labor Relations, Cornell University.

strategies involving the preselection of "predictable" subgroups from a sample of recruits.

## METHOD

### Subjects

Two hundred and four customer engineers (CEs) working for a large midwestern computer manufacturer participated in the study. The subjects were selected from five job grades, having a mean age of 28.8 years ( $SD=4.85$ ). The average subject had been a CE for 34.9 months. All subjects were male, and 96% were white.

### Predictor Variables

Tests used in the study and investigated as possible predictors (or moderators) were the following:

1. SRA Adaptability Test
2. Harris Inspection Test (HIT)
3. Employee Aptitude Survey (EAS)
  - Visual Pursuit
  - Visual Speed and Accuracy
  - Space Visualization
  - Numerical Reasoning
  - Symbolic Reasoning
4. Ghiselli Self-Description Inventory (SDI)
  - Supervisory Ability
  - Intelligence
  - Initiative
  - Self Assurance
  - Decisiveness
  - Masculinity-Femininity
  - Maturity
  - Working Class Affinity
  - Need for Achievement
  - Need for Self-Actualization
  - Need for Power
  - Need for High Financial Reward
  - Need for Security

### The Criterion

The criterion variable was an overall ranking of the CEs, based on a general appraisal of their job proficiency. Scores ranged from 1 to 9 ( $M=4.80$ ;  $SD=2.13$ ).

Data on 13 other variables, including demographic information and training data on the subjects, as well as demographic information on the rating supervisors, were also gathered.

### Procedure

The sample of 204 subjects was randomly subgrouped into a development group and a holdout group, composed of 131 and 73 subjects, respectively. Multiple linear regression was employed in a stepwise manner to predict the decile rank criterion using the SRA Adaptability Test, the HIT, and all of the EAS and SDI scales. Using the development group, a battery of three tests—the EAS Space Visualization Test, and the Initiative and Need for Security scales of the SDI—was

identified as the most predictive combination of tests, meeting the  $F$  test of significance at the 1% level of confidence. The multiple  $R$  was cross-validated on the holdout sample ( $N=73$ ), and the shrunken  $r$  was used to compute predicted test battery scores in the development group ( $N=131$ ). These predicted battery scores, as well as the corresponding criterion values, were transformed to  $Z$  scores. Then, using the method described by Ghiselli (1956), the coefficient of unpredictability,  $D$ , was calculated for these subjects.<sup>3</sup>

In an attempt to find a moderator variable, that is, a predictor of predictability, all other test data and training information were correlated with the  $D$  statistic in the development group. The average tenure of the CE on his last three jobs was found to be the best predictor of  $D$ . The relationship was cross-validated on the holdout group. Then, using the equation of the form

$$D = a + b(T)$$

where  $D$  = unpredictability and  $T$  = average tenure on last three jobs, seven arbitrarily selected values of  $D$  were used to determine seven critical (maximum) values of average tenure ( $T$ ) that would serve to determine subgroups of increasingly unpredictable subjects. The  $D$  values used were .75, .80, .85, .90, .95, 1.00, and 1.05. Subjects in the development group were then sorted into groups according to their predicted unpredictability status, using their tenure scores, and seven linear multiple regressions were run on increasingly larger (and more unpredictable) subgroups of the development group. Where the sizes of the corresponding groups in the holdout sample were not too small, these multiple  $R$ s were cross-validated on the holdout sample.

The Tilton overlap statistic ( $O$ ) was calculated for the distributions of criterion scores of the subjects included in each regression analysis as compared to those excluded as the result of their high tenure scores. This check was made in order to monitor any sampling differences in terms of criterion levels that would potentially affect the increase in multiple  $R$  at any stage, due to restriction of range on the performance variable. Further, using this statistic, we were able to investigate whether subjects of any particular criterion range were more likely to be discarded as unpredictable than others. Finally, the standard error of estimate associated with each subsample stage and each multiple  $R$  was calculated. Increases in the cross-validated  $r$  value and decreases in the corresponding accuracy of estimate were seen as benefits resulting from the application of the moderator, while the decreases in sample size needed to attain these benefits were seen as "costs."

In order to provide an estimate of the degree to which a hypothetical, perfect predictor of predictability would affect gains in  $R^2$  and  $\sigma_{es}$  for the same relative losses in sample size, the test battery was applied to seven subgroups of subjects whose actual  $D$  scores were the smallest. Thus, a "ceiling" estimate of the maximum possible  $R$  and minimum possible  $\sigma_{es}$  was calculated to correspond to each of the first seven prediction strat-

<sup>3</sup>  $D$  = Absolute difference between the standardized predictor score and the standardized criterion score for each subject, i.e.,  $(D = |Z_p - Z_c|)$ .

TABLE 1

SAMPLE SIZES, IMPROVED  $R^2$  AND  $\sigma_{est}$  FOR RUNS 1 THROUGH 7 USING TENURE AS MODERATOR VARIABLE ON DEVELOPMENT GROUP ( $N = 131$ )

Run number	$D$	Maximum tenure score in months	Sample <sup>a</sup> (in percent)	$R$	$R^2$	$\sigma_{est}$
1	.75	13	26	.524	.275	1.67
2	.80	19	38	.459	.211	1.82
3	.85	35	65	.452	.204	1.95
4	.90	51	79	.438	.192	1.89
5	.95	68	85	.444	.197	1.85
6	1.00	86	90	.413	.171	1.88
7	1.05	108	92	.407	.166	1.89
Total group			100	.395	.156	1.96

<sup>a</sup> The sample sizes are cumulative, and thus the subjects in each run number are subsumed in the subgroups at all other higher numbered runs.

egies, wherein the imperfect moderator—tenure—served as the basis for inclusion or exclusion of subjects. Again, multiple  $R$ s, Tilton  $O$ s, and  $\sigma_{est}$  were calculated at each state.

Finally, to demonstrate the *applied utility* of the moderator variable, a hypothetical selection exercise was conducted in which the selection ratio was varied from .10 to .90 in the development group, and the mean and standard deviation criterion scores of the subjects hired under each condition using the real, empirical moderator, as well as the hypothetical, or perfect moderator, were compared to the corresponding statistics among subjects hired without the use of the moderator (i.e., on the basis of their total test battery scores alone).

For example, given the task of hiring 13 CEs from the 131 subjects in the development group (selection ratio = .10), the means and SDs on the performance criterion of those 13 subjects whose test battery totals were the highest were compared to (a) the corresponding figures of the 13 subjects who were preselected according to their status on the tenure variable and then selected on the basis of their battery scores and (b) the same figures among those subjects who were preselected on the basis of their actual  $D$  scores, and then selected with the use of the battery.

## RESULTS

Using the entire development group ( $N = 131$ ), the multiple  $R$  calculated between the criterion and the test battery was .395 ( $\sigma_{est} = 1.96$ ). When cross-validated on the complete holdout group ( $N = 73$ ),  $r$  shrunk to .298 ( $\sigma_{est} = 2.08$ ).

The correlation coefficient calculated between Ghisell's  $D$  statistic and the best predictor of predictability (average tenure on last three jobs) was .295. When cross-validated on the holdout group,  $r$  fell to .20.

The relationship between tenure and predictability was a negative one, with less tenure being associated with greater predictability.

The critical tenure scores and the number of subjects in the development group with tenure scores below each critical value appear in Table 1. For instance, using the cross-validated relationship between  $D$  and tenure, and setting  $D$  equal to .75, we find that the maximum tenure score acceptable to select this first, most predictable group, was 13 months. Only 34 of the 131 subjects in the group had tenure scores below 13. For these 34 CEs, the multiple  $R$  calculated between the criterion and the same three-test battery was .524 ( $R^2 = .275$ ;  $\sigma_{est} = 1.67$ ). In this manner, the seven increasingly more unpredictable subgroups were used to compute multiple  $R$ s and standard errors of estimate. Table 1 summarizes these first seven runs.

As shown in Table 2, all of the multiple  $r$ s found in runs 1 through 7 shrunk, and the corresponding  $\sigma_{est}$  figures were all larger in the cross-validation group than in the development group. Because the number of subjects in the holdout group with tenure scores below the critical values for the first two runs was too small, the  $R$ s of .524 and .459 could not be cross-validated. Further, the same number of subjects ( $N = 60$ ) had the tenure scores necessary to cross-validate both runs 5 and 6, so their cross-validated  $r$ s and standard errors of estimate were the same.

By using the CEs actual  $D$  scores as the basis

TABLE 2

SAMPLE SIZES, CROSS-VALIDATED  $r^2$  VALUES FOR RUNS 3 THROUGH 7 AND  $\sigma_{est}$  USING TENURE AS MODERATOR VARIABLE ON HOLDOUT GROUP ( $N = 73$ )

Run number	$D$	Maximum tenure score in months	Sample <sup>a</sup> (in percent)	$r$	$r^2$	$\sigma_{est}$
3 CV	.85	35	59	.402	.162	1.85
4 CV	.90	51	75	.382	.146	1.91
5 CV	.95	68	82	.301	.091	2.02
6 CV	1.00	86	82	.301	.091	2.02
7 CV	1.05	108	89	.271	.073	2.08
Total group			100	.298	.089	2.08

<sup>a</sup> The sample sizes are cumulative, and thus the subjects in each run number are subsumed in the subgroups at all other higher numbered runs.

for computing multiple  $R$ s (again using the same test battery) we computed an estimate of the "ceiling"  $R$ , which could be attained if tenure had correlated perfectly with  $D$ . Thus, in runs 8 through 14, the same numbers of subjects were selected for the development of  $R$  as in runs 1 through 7, but in these runs the actual, most predictable subgroups of the development sample were used. Thus, run 8 used 34 subjects ( $R = .978$ ;  $\sigma_{est} = .434$ ), and run 14 used 121 subjects. Table 3 presents the results of runs 8 through 14, and Table 4 presents the corresponding cross-validated values, as calculated on the corresponding subgroups of the holdout sample. Because of small sample sizes, runs 8 and 9 could not be cross-validated.

The Tilton overlap statistics calculated for the distributions of the criterion scores of the subgroups included in each run, as compared to those of the subjects excluded in each run, suggested that the subgrouping procedure extracted CEs of equal mean criterion ratings, but with different variabilities, as would be predicted. The median percentage overlap for the 22 runs was 93%, the range was from 77% to 100%. In other words, unpredictable subjects were discarded from both sides of the regression line at each pass. Table 5 presents the means and standard deviations of the criterion scores for the "included" and "excluded" subjects at each pass, as well as the numerical differences between the various pairs of sigmas. As we would expect, the differences between the sigmas were greater in runs 8 through 14

TABLE 3

SAMPLE SIZES, IMPROVED  $R^2$  AND  $\sigma_{est}$  FOR RUNS 8 THROUGH 14 USING ACTUAL  $D$  SCORE AS MODERATOR ON DEVELOPMENT GROUP ( $N = 131$ )

Run number	Sample <sup>a</sup> (in percent)	$R$	$R^2$	$\sigma_{est}$
8	26	.978	.956	.434
9	38	.949	.901	.442
10	65	.803	.645	1.14
11	79	.643	.413	1.17
12	85	.552	.305	1.67
13	90	.472	.223	1.81
14	92	.407	.165	1.90
Total group	100	.395	1.56	1.96

<sup>a</sup> The sample sizes are cumulative, and thus the subjects in each run number are subsumed in the subgroups at all other higher numbered runs.

TABLE 4

SAMPLE SIZES, CROSS-VALIDATED  $r^2$  VALUES FOR RUNS 10 THROUGH 14 AND  $\sigma_{est}$  USING ACTUAL  $D$  SCORE AS MODERATOR VARIABLE ON HOLDOUT GROUP ( $N = 73$ )

Run number	Sample <sup>a</sup> (in percent)	$r$	$r^2$	$\sigma_{est}$
10 CV	73	.749	.561	1.31
11 CV	79	.692	.479	1.42
12 CV	90	.584	.341	1.64
13 CV	93	.528	.279	1.79
14 CV	100	.298	.089	2.08
Total group	100	.298	.089	2.08

<sup>a</sup> The sample sizes are cumulative, and thus the subjects in each run number are subsumed in the subgroups at all other higher numbered runs.

TABLE 5

MEANS AND STANDARD DEVIATIONS OF CRITERION SCORES OF SUBJECTS DISCARDED AND ACCEPTED FOR PREDICTION BY THE MODERATOR VARIABLE

Using tenure as moderator						Using <i>D</i> as moderator					
Run	<i>N</i>	Statistic	Unpredictable	Predictable	Difference	Run	<i>N</i>	Statistic	Unpredictable	Predictable	Difference
1	34	<i>M</i>	4.67	5.18	-.51	8	34	<i>M</i>	4.79	4.82	-.03
		<i>SD</i>	2.18	1.96	.22			<i>SD</i>	2.16	2.07	.09
2	50	<i>M</i>	4.85	4.72	.13	9	50	<i>M</i>	4.83	4.76	.07
		<i>SD</i>	2.19	2.05	.14			<i>SD</i>	2.22	1.96	.26
3	85	<i>M</i>	4.57	4.93	-.36	10	85	<i>M</i>	4.47	4.98	-.51
		<i>SD</i>	2.02	2.19	-.17			<i>SD</i>	2.47	1.91	.56
4	103	<i>M</i>	4.71	4.83	-.12	11	103	<i>M</i>	4.43	4.90	-.47
		<i>SD</i>	2.26	2.10	.16			<i>SD</i>	2.67	1.96	.71
5	112	<i>M</i>	4.79	4.80	-.01	12	112	<i>M</i>	4.21	4.90	-.69
		<i>SD</i>	2.51	2.07	.44			<i>SD</i>	2.76	2.00	.76
6	118	<i>M</i>	4.38	4.85	-.47	13	118	<i>M</i>	4.31	4.86	-.55
		<i>SD</i>	2.72	2.06	.66			<i>SD</i>	2.78	2.05	.73
7	121	<i>M</i>	4.10	4.86	-.76	14	121	<i>M</i>	4.10	4.86	-.76
		<i>SD</i>	2.69	2.08	.61			<i>SD</i>	2.69	2.08	.61
CV3	43	<i>M</i>	5.47	5.09	.38	CV10	53	<i>M</i>	5.10	5.30	-.20
		<i>SD</i>	2.40	2.02	.38			<i>SD</i>	2.69	1.98	.71
4	55	<i>M</i>	5.11	5.29	-.18	11	58	<i>M</i>	5.40	5.21	.19
		<i>SD</i>	2.54	2.07	.47			<i>SD</i>	2.95	1.96	.99
5	60	<i>M</i>	5.84	5.12	.72	12	66	<i>M</i>	5.00	5.27	-.27
		<i>SD</i>	2.41	2.12	.29			<i>SD</i>	3.56	2.02	1.54
6	60	<i>M</i>	5.84	5.12	.72	13	68	<i>M</i>	5.00	5.26	-.26
		<i>SD</i>	2.41	2.12	.29			<i>SD</i>	3.32	2.11	1.21
7	65	<i>M</i>	6.37	5.11	1.26	14	73	<i>M</i>	—	5.25	
		<i>SD</i>	2.13	2.16	-.03			<i>SD</i>	—	2.18	

and runs 10 CV through 14 CV, where the most unpredictable people were in fact those excluded at each stage.

Table 6 presents the means and standard deviations of the subjects having the highest predictor scores in each moderator group when subjects' actual *D* scores were used to preselect CEs. A trade-off in value in terms of these criterion scores appears between the use of the moderator on the one hand and the familiar selection ratio effect on the other.

Table 7 presents the corresponding criterion data that resulted from the use of the CEs tenure scores as the moderator variable.

#### DISCUSSION

The present data have demonstrated several principles in moderated selection strategies. First, through the use of the predictor of predictability model,  $R^2$  increases and  $\sigma_{\text{ext}}$  decreases, even when the moderator can

account for as little as 4% of the variance in *D*. Table 2 shows that by decreasing the sample size from 73 to 43 in the holdout group,  $r^2$  increased from .089 to .162, and  $\sigma_{\text{ext}}$  decreased from 2.08 to 1.85.

Whether predictors of predictability can be found in practical settings that are more valid than the tenure variable used here is doubtful, given that the moderator itself does not serve in the multiple regression model as a predictor of job performance. In the present study, an extra stepwise regression was run using all training and demographic variables as possible predictors, in order to determine whether the moderator, tenure, would itself serve as a predictor in the test battery. Tenure was not found to be a valid predictor of the criterion when used in combination with the other variables. This check was conducted in light of the criticism raised by Zedeck (1971), that many so-called "moderator" variables are in

TABLE 6  
MEAN CRITERION SCORES OF THE CEs HAVING THE HIGHEST PREDICTOR SCORES  
IN EACH MODERATOR (*D*) SUBGROUP

Selection ratio	<i>N</i>	Size of moderated group						Complete group ( <i>N</i> = 131)
		34	50	85	103	112	118	
.10	13	6.84 (1.14)	7.15 (1.07)	7.23 (1.01)	7.23 (1.01)	7.23 (1.01)	6.92 (1.32)	6.92 (1.32)
.20	26	5.42 (1.86)	6.38 (1.27)	6.73 (1.22)	6.77 (1.21)	6.54 (1.50)	6.42 (1.58)	6.04 (1.75)
.30	39	5.33 (1.84)	5.62 (3.35)	6.18 (1.57)	6.18 (1.71)	6.13 (1.81)	6.15 (1.76)	5.90 (1.88)
.40	52		5.54 (2.96)	6.08 (1.56)	6.12 (1.66)	5.98 (1.78)	5.85 (1.86)	5.58 (2.08)
.50	65		5.11 (2.82)	5.82 (1.72)	5.92 (1.75)	5.94 (1.81)	5.78 (1.96)	5.49 (2.05)
.60	78		4.74 (2.71)	5.53 (1.80)	5.69 (1.83)	5.68 (1.92)	5.60 (1.98)	5.45 (2.11)
.70	91			5.16 (1.90)	5.40 (1.92)	5.45 (1.96)	5.43 (2.00)	5.34 (2.08)
.80	104				5.08 (2.00)	5.16 (2.02)	5.14 (1.94)	5.17 (2.09)
.90	117					4.94 (2.03)	4.94 (2.02)	4.97 (2.11)

Note. SDs are in the parentheses.

TABLE 7  
MEAN CRITERION SCORES OF THE CEs HAVING THE HIGHEST PREDICTOR SCORES  
IN EACH MODERATOR (TENURE) SUBGROUP

Selection ratio	<i>N</i>	Size of moderated group						Complete group ( <i>N</i> = 131)
		34	50	85	103	112	118	
.10	13	6.23 (1.64)	6.08 (1.93)	6.54 (1.90)	6.54 (1.90)	6.38 (1.85)	6.46 (1.85)	6.46 (1.85)
.20	26	5.42 (2.00)	5.23 (2.21)	6.27 (1.91)	6.04 (1.87)	6.04 (1.87)	5.92 (1.79)	5.92 (1.79)
.30	39		4.90 (2.15)	5.77 (2.05)	5.64 (2.05)	5.59 (2.06)	5.82 (1.92)	5.82 (1.92)
.40	52			5.46 (2.27)	5.44 (2.14)	5.52 (1.99)	5.60 (1.90)	5.57 (1.98)
.50	65			5.34 (2.23)	5.35 (2.16)	5.42 (2.12)	5.46 (2.06)	5.48 (2.11)
.60	78			5.06 (2.33)	5.26 (2.13)	5.33 (2.07)	5.32 (2.06)	5.36 (2.10)
.70	91				5.02 (2.12)	5.14 (2.08)	5.21 (2.06)	5.27 (2.09)
.80	104					4.90 (2.08)	5.00 (2.07)	5.11 (2.12)
.90	117						4.85 (2.07)	4.94 (2.12)

Note. SDs are in the parentheses.

fact not really moderators, since their inclusion in the multiple regression equation would assist in explaining criterion variance. This was not the case.

A second observation of interest is the fact that unpredictable subjects were found at all ranges of the criterion variable. Unpredictable recruits who are discarded before selection may be either high- or low-potential performers, as shown in Table 5. Whether or not the discarding of potentially competent employees actually constitutes a loss to an organization will be determined by the costs of recruiting, the proportion of the predictable applicants who are found to be acceptable, and the cost of allowing competent individuals to find employment with competitors. In the case of a "tight" labor market for managerial, professional, or technical personnel, this cost could be considerable, and should be a factor in the decision of whether or not to use a moderated selection strategy.

The present study also demonstrates that moderator variables can be found, and that they can withstand the test of cross-validation without shrinking and losing all of their value.

Further, through the use of actual  $D$  scores in runs 8 through 14 (as shown in Tables 3 and 4), it was demonstrated how a highly valid moderator (here a hypothetical, perfect predictor of predictability) could serve to further enhance our utility in selection, through greater increases in  $R^2$  per percentage decrease in  $N$ , and through greater reduction in errors of estimate. In the cross-validation group, for example, we were able to increase  $r^2$  from .298 to .749 and reduce  $\sigma_{est}$  from 2.08 to 1.31 by discarding the 20 most unpredictable subjects from our group of 73. (This method would of course be impossible in a real selection situation, since no criterion information would be available for applicants.)

The means and standard deviations shown for the criterion scores of the subjects included in each stage of the analysis, as compared to those excluded at each stage, provide a numerical illustration of how moderator variables work to identify cases that fall away from the regression line, in order that we may exclude them and proceed to make predictions for only those cases which are more predictable.

As shown in Table 5, when tenure was the moderator the differences between the standard deviation of the "rejected" and the "predictable" subjects on the criterion variable were positive in 9 of the 11 cases, indicating that the most unpredictable cases were almost always being discarded through the application of the moderator. When the actual unpredictability score,  $D$ , was used as the moderator, the standard deviation of the criterion scores of the unpredictables was always greater than that of the predictables, indicating that the perfect moderator did a better job of prescreening unpredictable recruits than did the tenure variable.

The hypothetical selection problems summarized in Tables 6 and 7 provide a numerical illustration of how effective the moderators are in helping us to select a given number of employees from a group of applicants.

In Table 6 we can see two forces at work determining the degree to which the perfect moderator assists in selecting successful employees. Reading across the rows of Table 6, thus holding the number of CEs to be hired constant, and selecting this number from increasingly larger subgroups (which in effect means decreasing the selection ratio), mean performance scores increased steadily and then decreased as we approached the full sample size ( $N = 131$ ). In using the larger subgroups, we were able to increase the average criterion performance, since at each stage there were more subjects from which to select the 13 top performers. Thus there was the familiar selection ratio effect wherein we are able to choose better employees through using a larger pool of applicants. However, as we added more and more people to the applicant pool, we added increasingly more unpredictable subjects, and the result was that some of the 13 predicted to be top performers were in fact overpredicted. Hence the mean criterion scores fell again. And as we moved to subgroups of more unpredictable subjects, the criterion variance increased again because of the wide oval shape of the scatterplot relating the criterion and the predictor battery.

The same general patterns appeared in Table 7, but not as clearly since the tenure moderator was not doing the perfect job of

prescreening applicants that we accomplished in Table 6 through the use of actual  $D$  scores.

When our moderator was perfectly accurate in identifying predictable subjects, it was useful in assisting us in hiring more effective CEs than we were able to hire without the moderator (see Table 6). As the selection ratio increased (as we hired greater numbers of CEs), the optimal number of applicants to be tested increased. As the selection ratio approached 1 (where  $N = 104$  and greater), the moderators did not assist us at all. In fact, the mean performance ratings without the moderator were greater than those that resulted when either tenure or  $D$  was used.

When we compare the results in Table 7 (where tenure was the moderator) to those of Table 6, we can assess the practical value of our empirically identified moderator and compare it to the utility derived by using the perfect moderator. Reading across the rows of Table 7 reveals that the same sort of trade-off relationship between the selection ratio effect on the one hand and the moderator effect on the other was occurring. However, to the extent that tenure was not a perfectly accurate predictor of predictability, this moderator was less useful than in those cases where  $D$  was used. By using subgroups of the total sample, we lost the benefit of the selection ratio effect without receiving in return sufficient accuracy in identifying predictable high performers. Consequently, of the nine rows in Table 7, the moderator allowed us to improve the mean and variance in only three cases (where the selection ratio equals .10, .20, and .40) above that in the total group. In the other cases, we would have been able to select a better performing and more homogeneous group of CEs through using the entire sample without the tenure moderator, thus taking full advantage of the selection ratio effect.

It can be concluded that our empirically found moderator was of little or no benefit in helping to select given numbers of CEs. On the other hand,  $D$ , the perfect moderator, was

effective in this task. Presumably, the more accurate the moderator variable, the more useful it will be in such practical applications. Moreover, to the extent to which it would be difficult to find a moderator variable in actual selection settings which is more accurate than that found here, it is questionable whether moderated selection strategies with high selection ratios will be of any value in selecting the top performers from our applicant pool.

### CONCLUSIONS

Moderated selection strategies involving the preselection of recruits based on the likelihood that later statistical prediction will be accurate for them can be useful in increasing predictive accuracy. However, as demonstrated here, there are costs, some more obvious than others, related to the use of moderators. The validity of the predictor without the moderator, the accuracy of the moderator itself, the costs of false positives and false negatives, the selection ratio, and the cost of using the moderating test are some of the major factors that should be considered before the decision is made to employ the moderator. As these conditions vary, so will the net incremental utility derived from the use of the moderator variable in the personnel selection process.

### REFERENCES

- GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, 1956, 40, 374-378.
- HOBERT, R. D. Moderating effects in the prediction of managerial scores from psychological test scores and biographical factors. Unpublished doctoral dissertation, University of Minnesota, 1965.
- HOBERT, R. D., & DUNNETTE, M. D. Development of moderator variables to enhance the prediction of managerial effectiveness. *Journal of Applied Psychology*, 1967, 51, 50-64.
- MCNEMAR, Q. Moderation of a moderator technique. *Journal of Applied Psychology*, 1969, 53, 69-72.
- ZEDECK, S. Problems with the use of moderator variables. *Psychological Bulletin*, 1971, 76, 295-310.

(Received December 13, 1971)

## A METHOD FOR EVALUATING ALTERNATIVE RECRUITING- SELECTION STRATEGIES:

### THE CAPER MODEL<sup>1</sup>

WILLIAM A. SANDS<sup>2</sup>

*Naval Personnel Research and Development Laboratory, Washington, D.C.*

Managers of personnel systems justifiably demand an estimate of the payoff, in dollars, which can be expected to result from the implementation of a proposed selection program. The Cost of Attaining Personnel Requirements (CAPER) model determines an optimal recruiting-selection strategy. Specifically, the CAPER model provides the personnel manager with the information necessary to minimize the estimated total cost of recruiting, selecting, inducting, and training a sufficient number of persons to meet a specified quota of satisfactory personnel. This article describes the CAPER model and illustrates the application of the model to a personnel recruiting-selection problem. The advantages and limitations of the model are discussed.

Managers of personnel systems justifiably demand an estimate of the payoff, in dollars, which can be expected to result from the implementation of a proposed selection program.

Taylor and Russell (1939) published a series of tables which illustrated that the value of a test was a function of three considerations: (a) the validity coefficient (the predictor-criterion correlation); (b) the base rate (the proportion of persons currently accepted who are satisfactory); and (c) the selection ratio (the proportion of applicants accepted). These tables show that, for a fixed validity coefficient and base rate, lowering the selection ratio will improve the success ratio (the proportion of persons accepted who are satisfactory).

It might appear that, for a given test validity and base rate, a personnel manager always should lower the selection ratio (i.e., become more selective). However, the optimal strategy is not this simple. If, as is often the case, a quota exists for a certain number of satisfactory personnel, lowering the selection ratio forces the recruiting and selection effort to be expanded. This strategy may or may not be cost effective.

The purpose of this article is to demonstrate

the Cost of Attaining Personnel Requirements (CAPER) model. This model is designed to evaluate the cost consequences of alternative recruiting-selection strategies. Specifically, the CAPER model determines an optimal recruiting-selection strategy for minimizing the estimated total cost of recruiting, selecting, inducting, and training a sufficient number of persons to meet a specified quota of satisfactory personnel. In addition, the CAPER model considers the cost of an erroneous acceptance (selecting a person for a training program who subsequently fails to graduate) and the cost of an erroneous rejection (rejecting a person who would have succeeded if given the opportunity).<sup>3</sup> Readers interested in similar approaches are referred to Doppelt and Bennett (1953) and Dunnette (1966).

### METHOD

The personnel manager desiring to utilize the CAPER model to aid in the formulation of an optimal recruiting-selection policy must be able to specify the following information: the quota, the base rate, and the proportion of previous graduates and failures (separately) who would have qualified for acceptance at each possible cutting score on the new test (if it had been used for selection).<sup>4</sup> In addition, the following cost data per person must be specified: recruiting, selection, induction

<sup>1</sup> A brief, management-oriented paper describing this model was presented at the 12th Annual Military Testing Association (MTA) Conference (Sands, 1970).

<sup>2</sup> The opinions expressed are those of the author and do not necessarily reflect those of the Navy Department.

Requests for reprints should be sent to William A. Sands, Research Psychologist, Personnel Measurement Research Division, Naval Personnel Research and Development Laboratory, Building 200, Washington Navy Yard, Washington, D.C. 20374.

<sup>3</sup> This terminology follows Curtis (1967) and seems far more appropriate to personnel selection than the more traditional terminology based upon a disease model of medicine (e.g., "false positives").

<sup>4</sup> Under the assumption of a bivariate normal distribution, these proportions could be estimated from the usual test-criterion statistics (means, standard deviations, and correlation).

TABLE 1  
PROPORTION OF GRADUATES AND FAILURES QUALIFIED FOR ACCEPTANCE  
USING VARIOUS CUTTING SCORES ON THE TEST

Cutting score	Graduates			Failures		
	Frequency	Qualified for acceptance		Frequency	Qualified for acceptance	
		Number	Proportion <sup>a</sup>		Number	Proportion <sup>a</sup>
0	0	500	1.000	0	500	1.000
1	0	500	1.000	7	500	1.000
2	0	500	1.000	10	493	.986
3	0	500	1.000	17	483	.966
4	2	500	1.000	29	466	.932
5	6	498	.996	45	437	.874
6	11	492	.984	58	392	.784
7	21	481	.962	68	334	.668
8	36	460	.920	72	266	.532
9	52	424	.848	66	194	.388
10	66	372	.744	52	128	.256
11	72	306	.612	36	76	.152
12	68	234	.468	21	40	.080
13	58	166	.332	11	19	.038
14	45	108	.216	6	8	.016
15	29	63	.126	2	2	.004
16	17	34	.068	0	0	.000
17	10	17	.034	0	0	.000
18	7	7	.014	0	0	.000

Note. These data are based on an article by McNemar (1969).

<sup>a</sup> These proportions are used for the CAPER model.

(processing), training, erroneous acceptance, and erroneous rejection.

A hypothetical personnel recruiting-selection problem will be used to illustrate the application of the model equations. The manager of the personnel program needs 50 graduates at the end of the next training period. The only admission requirement under the ordinary selection procedure, is a medical clearance. Data on an experimental aptitude test are available for a large random sample of applicants previously admitted to the program. This test was not used in the ordinary selection procedure, and students' scores were not available to the instructor. Table 1 shows that 500 of the 1,000 persons graduated, indicating a .500 base rate.

The best estimate for the cost of recruiting an individual is \$50.<sup>b</sup> This estimate covers the salaries of the recruiting personnel and advertising expenses.

<sup>b</sup> The assumption of linear costs is questionable in most cases. For example, the cost of recruiting 200 persons typically is more than double the cost of recruiting 100 persons. Unfortunately, many (if not most) organizations do not maintain detailed personnel cost records to enable them to specify, for example, that recruiting the first 100 individuals costs \$50 each, while recruiting the second 100 persons costs \$60 each. This would reflect the increasing difficulty of contacting more and more applicants. A version of the CAPER model that will handle stepwise-linear costs is reported elsewhere (Sands, 1971b).

The cost of the ordinary selection procedure (medical examination) is \$20 including the physicians' salaries and laboratory fees. The cost of administering and scoring the experimental test is \$5. The induction cost per individual is estimated as \$15 and includes the administrative and clerical expenses involved in processing the selectee into the training program. The cost estimate for training is \$400 and covers the salary of the instructor and course materials. The cost of accepting a person into the training program who subsequently fails to graduate (erroneous acceptance) is set at \$100 and includes the administrative costs of termination, an estimate of the loss incurred as a result of decreased morale of persons remaining in the program, and the cost of negative behavior (e.g., damage to training equipment). The cost of rejecting an individual who would have graduated if he had been accepted (erroneous rejection) is estimated as \$80. This cost estimate reflects the manager's opinion of the disadvantage to the program when a competing company correctly accepts the individual.<sup>c</sup>

<sup>c</sup> Accurate estimation of these various costs is difficult, particularly the costs of erroneous personnel decisions. For example, it could be argued that competition has no impact on a program and that an erroneous rejection involves no cost to the institution, providing the quota for graduates is met. In any case, it should be recognized that ignoring the costs of the two types of decision errors is equivalent to setting them equal to zero.

The first time the set of equations discussed below is employed, the results (numbers of persons and dollar costs) are computed for the ordinary selection procedure, in which the new test is not administered. Then the experimental selection procedure (the medical examination plus the test) is evaluated. The set of equations is used once for each possible cutting score on the new test. Finally, the optimal use of the experimental selection procedure is compared to the ordinary selection procedure in terms of the estimated total cost of meeting the quota of satisfactory personnel. A cutting score of eight on the test will be used for illustration.

Equation 1 gives the formula for estimating the number of applicants who must be recruited in order to meet the quota:

$$NR = Q / [(BR)(PG_i)] \quad BR > 0; \quad PG_i > 0 \quad [1]$$

where NR is the number recruited, Q is the quota of satisfactory personnel, BR is the base rate, and PG<sub>i</sub> is the proportion of graduates who would qualify for acceptance at the *i*th cutting score on the test.<sup>7</sup>

Substituting the data pertinent to the ordinary selection procedure into Equation 1 gives:

$$\begin{aligned} NR_0 &= 50 / [(0.500)(1.000)] \\ &= 100 \end{aligned}$$

Similarly, for the experimental selection procedure (*i* = 8):

$$\begin{aligned} NR_e &= 50 / [(0.500)(0.920)] \\ &= 109 \end{aligned}$$

Equation 2 gives the formula for estimating the number of erroneous acceptances:

$$NEA = (NR)(1 - BR)(PF_i) \quad [2]$$

where NEA is the number of erroneous acceptances, PF<sub>i</sub> is the proportion of failures who would qualify for acceptance at the *i*th cutting score on the test, and the remaining symbols are defined above.

Substituting the data pertinent to the ordinary selection procedure into Equation 2 gives:

$$\begin{aligned} NEA_0 &= (100)(1.000 - 0.500)(1.000) \\ &= 50 \end{aligned}$$

Similarly, for the experimental selection procedure (*i* = 8):

$$\begin{aligned} NEA_e &= (109)(1.000 - 0.500)(0.532) \\ &= 29 \end{aligned}$$

Equation 3 gives the formula for estimating the number of erroneous rejections:

$$NER = (NR)(BR)(1 - PG_i) \quad [3]$$

where NER is the number of erroneous rejections, and the remaining symbols are defined above.

Substituting the data pertinent to the ordinary selection procedure into Equation 3 gives:

$$\begin{aligned} NER_0 &= (100)(0.500)(1.000 - 1.000) \\ &= 0 \end{aligned}$$

<sup>7</sup> Inasmuch as the test would not be administered under the ordinary selection procedure, no graduate or failure would be rejected on the basis of the test score and, therefore, PG = PF = 1.000 for this procedure.

Similarly, for the experimental selection procedure (*i* = 8):

$$\begin{aligned} NER_e &= (109)(0.500)(1.000 - 0.920) \\ &= 4 \end{aligned}$$

Equation 4 gives the formula for estimating the number of persons who will be accepted:

$$NA = Q + NEA \quad [4]$$

where NA is the number accepted, and the remaining symbols are defined above.

Substituting the data pertinent to the ordinary selection procedure into Equation 4 gives:

$$\begin{aligned} NA_0 &= 50 + 50 \\ &= 100 \end{aligned}$$

Similarly, for the experimental selection procedure (*i* = 8):

$$\begin{aligned} NA_e &= 50 + 29 \\ &= 79 \end{aligned}$$

Unlike the above four equations that were applicable to both the ordinary selection procedure and the experimental selection procedure, the estimation of total cost requires separate equations.

Equation 5a gives the formula for estimating the total cost of employing the ordinary selection procedure to meet the quota of satisfactory personnel:

$$\begin{aligned} TC_0 &= [(NR)(CR)] + [(NR)(CO)] + [(NA)(CI)] \\ &\quad + [(NA)(CT)] + \{[(NEA)(CEA)] \\ &\quad + \{(NER)(CER)\}\} \quad [5a] \end{aligned}$$

where TC<sub>0</sub> is the total cost of using the ordinary selection procedure to meet the quota, CR is the cost of recruiting a person, CO is the cost of administering the ordinary selection procedure, CI is the cost of inducting a person, CT is the cost of training a person, CEA is the cost of an erroneous acceptance, CER is the cost of an erroneous rejection, and the remaining symbols are defined above.

Substituting the data pertinent to the ordinary selection procedure into Equation 5a gives:

$$\begin{aligned} TC_0 &= [(100)(\$50)] + [(100)(\$20)] + [(100)(\$15)] \\ &\quad + [(100)(\$400)] + \{[(50)(\$100)] \\ &\quad + [(0)(\$80)]\} \\ &= [\$5,000] + [\$2,000] + [\$1,500] \\ &\quad + [\$40,000] + [\$5,000] \\ &= \$53,500 \end{aligned}$$

Equation 5b gives the formula for estimating the total cost of employing the experimental selection procedure to meet the quota:

$$\begin{aligned} TC_e &= [(NR)(CR)] + \{[(NR)(CO)] + [(NR)(CE)]\} \\ &\quad + [(NA)(CI)] + [(NA)(CT)] \\ &\quad + \{[(NEA)(CEA)] + [(NER)(CER)]\} \quad [5b] \end{aligned}$$

where TC<sub>e</sub> is the total cost of using the experimental selection procedure, CE is the cost of administering the experimental test to a person, and the remaining symbols are defined above.

Using Equation 5b for the experimental selection procedure (*i* = 8) gives:

TABLE 2  
EXAMPLE OF CAPER MODEL OUTPUT DATA

Cut. score	No. rec.	No. acc.	No. err. acc.	No. err. rej.	Costs <sup>a</sup>						
					Recruit.	Select.	Induct.	Traing.	Err. dec.	Total	Grad.
NA <sup>b</sup>	100	100	50	0	5,000	2,000	1,500	40,000	5,000	53,500	1,070
0	100	100	50	0	5,000	2,500	1,500	40,000	5,000	54,000	1,080
1	100	100	50	0	5,000	2,500	1,500	40,000	5,000	54,000	1,080
2	100	99	49	0	5,000	2,500	1,485	39,600	4,900	53,485	1,070
3	100	98	48	0	5,000	2,500	1,470	39,200	4,800	52,970	1,059
4	100	97	47	0	5,000	2,500	1,455	38,800	4,700	52,455	1,049
5	100	94	44	0	5,000	2,500	1,410	37,600	4,400	50,910	1,018
6	102	90	40	1	5,100	2,550	1,350	36,000	4,080	49,080	982
7	104	85	35	2	5,200	2,600	1,275	34,000	3,660	46,735	935
8	109	79	29	4	5,450	2,725	1,185	31,600	3,220	44,180	884
9	118	73	23	9	5,900	2,950	1,095	29,200	3,020	42,165	843
10	134	67	17	17	6,700	3,350	1,005	26,800	3,060	40,915	818
11	163	62	12	32	8,150	4,075	930	24,800	3,760	41,715	834
12	214	59	9	57	10,700	5,350	885	23,600	5,460	45,995	920
13	301	56	6	101	15,050	7,525	840	22,400	8,680	54,495	1,090
14	463	54	4	181	23,150	11,575	810	21,600	14,880	72,015	1,440
15	794	52	2	347	39,700	19,850	780	20,800	27,960	109,090	2,182
16	1,471	50	0	685	73,550	36,775	750	20,000	54,800	185,875	3,718
17	2,941	50	0	1,421	147,050	73,525	750	20,000	113,680	355,005	7,100
18	7,143	50	0	3,521	357,150	178,575	750	20,000	281,680	838,155	16,763

Note. Abbreviations: Cut. score = cutting score; No. rec. = number received; No. acc. = number accepted; No. err. acc. = number of erroneous acceptances; No. err. rej. = number of erroneous rejections; Recruit. = recruitment; Select. = selection; Induct. = induction; Traing. = training; Err. dec. = erroneous decisions; Grad. = graduated.

<sup>a</sup> These costs are rounded to the nearest dollar.

<sup>b</sup> The information in this row pertains to the ordinary selection procedure and, therefore, a cutting score on the experimental variable is not applicable.

$$\begin{aligned}
 TC_a &= [(109)(\$50)] + [ \{ (109)(\$20) ] + [ (109)(\$5) ] \\
 &\quad + [(79)(\$15)] + [(79)(\$400)] + [ \{ (29)(\$100) ] \\
 &\quad + [ (4)(\$80) ] \\
 &= [\$5,450] + [\$2,725] + [\$1,185] \\
 &\quad + [\$31,600] + [\$3,220] \\
 &= \$44,180
 \end{aligned}$$

## RESULTS

The equations presented above yield five types of information: (a) number recruited, (b) number of erroneous acceptances, (c) number of erroneous rejections, (d) number accepted, and (e) total cost. These five estimates may suffice for many personnel program managers.

A more detailed list of information available from the model would include the five items listed above and a breakdown of total cost into five component parts: (a) recruiting, (b) selection, (c) induction, (d) training, and (e) erroneous decisions. Examination of Equations

5a and 5b reveals that these components of the total cost are contained in the five terms enclosed in brackets. Use of all available cost data provides a deeper insight into the consequences of altering the cutting score on the test.

Examination of Table 2 shows that as the manager becomes more selective (increases the cutting score), these consequences follow: (a) a greater number of persons must be recruited, (b) a smaller number of persons is accepted, (c) the number of erroneous acceptances decreases, and (d) there is an increase in erroneous rejections. These four consequences have cost implications. Recruiting and selection costs increase. These increases are accompanied by decreased induction and training costs. The cost of erroneous decisions, reflecting both erroneous acceptances and erroneous rejections, decreases at first, hits the cutting score that minimizes the sum of both costs ( $i = 9$ ) and then increases as the cutting score is raised farther. The most critical item,

total cost, likewise decreases to a point and subsequently increases.

The manager wants to achieve his quota of 50 graduates at a minimum total cost. Comparison of the estimated total cost of the ordinary selection procedure (\$53,500) with the experimental selection procedure using a cutting score of 8 (\$44,180) indicates that using the test operationally would be cost effective.

Perusal of the estimated total cost column of Table 2 shows that the minimum total cost of attaining the quota is \$40,915, using a cutting score of 10 on the test. The optimal recruiting-selection strategy is to recruit 134 persons. The best estimate is that 67 of these persons will qualify for acceptance. Seventeen of the selectees can be expected to fail the training course, leaving the 50 graduates required to meet the manager's quota. In comparison to the ordinary selection procedure, the optimal use of the experimental selection procedure will save an estimated \$12,585 or \$252 per graduate.

#### DISCUSSION

Implementation of the optimal recruiting-selection strategy should not be attempted in an automatic, mechanical fashion. The CAPER model is designed to provide useful planning information to managers of personnel systems, not to replace them, nor relieve them of the responsibility for sound decision making.

The results should be critically reviewed by all cognizant personnel to insure the feasibility of the optimal strategy; otherwise, serious problems can be encountered. For example, under certain input cost configurations, the optimal cutting score on the experimental variable (e.g., a test) may be so high that it is impossible to recruit a sufficient number of personnel, regardless of the financial resources invested. Although this type of problem can be avoided by the careful specification of cost estimates, it is difficult to foresee.

The most important advantage of the CAPER model from the standpoint of applied psychological research is the ease with which the results may be communicated. The results, numbers of persons and dollar costs, are readily understood by everyone, regardless of their interests or educational background.

The results of many other approaches—for example, the correlation model—are somewhat esoteric and often discourage rather than encourage communication (Thorndike, 1949). As Dunnette (1962) and Uhlaner (1960) point out, personnel research psychologists generally have focused their attention on selection and have ignored, or made only passing reference to, the fact that any selection strategy has implications for the entire personnel system. An optimal strategy for selection may not be optimal, or even feasible, for recruiting and/or training. When undue attention is paid to selection, to the exclusion of other components of the system, serious suboptimization of the entire personnel system can result. The CAPER model, unlike many alternative models, takes cognizance of the personnel system from initial recruitment through completion of training.

In the tradition of classical test theory, the correlation model focuses upon the accuracy of measurement. In contrast, the CAPER model is decision oriented and recognizes the necessity of taking into account the utility or cost of various decision-outcome combinations. Requiring an explicit estimate of the various types of costs decreases the likelihood that personnel policy decisions will be based on implicit, unrecognized, and frequently unwarranted assumptions (Cronbach & Gleser, 1965).

The flexibility of the CAPER model constitutes a major advantage for the personnel manager. The model allows for the separate specification of recruiting, selection, induction, training, and both types of erroneous decision costs. This enables the user to quickly and efficiently simulate the impact of various cost configurations for a particular problem. To facilitate this "gaming" use of the CAPER model, a user's manual including a FORTRAN computer program and detailed documentation has been prepared (Sands, 1971a). In addition, the personnel manager can adapt his recruiting-selection strategy readily to changes in quotas and/or alterations in the recruiting environment.

The model is quite general and could be used by the manager of many relatively large personnel systems. The simplicity of the mathematical approach would facilitate any modifications deemed necessary to "custom-

"tailor" the CAPER model for application to a particular personnel system.

Use of the CAPER model entails the assumption that all graduates of the training program are equally useful to the organization in terms of actual on-the-job performance. A further assumption required is that the predictor-criterion relationship is stable. This means that the base rate and the experimental variable frequency distributions for graduates and failures are based on a representative sample composed of a relatively large number of selectees. These assumptions (and often many others) are made by other models for personnel selection. However, it is worth mentioning that data based on a biased sample could yield seriously erroneous results.

Obviously, the utility of the CAPER model output data can be no better than the accuracy of the input data. If the cost estimates provided by the user are unrealistic, the cost forecasts and optimal recruiting-selection strategy will be misleading.

The CAPER model is an analytic rather than a stochastic model. This means that the output data are fixed by the values of the input data in a deterministic fashion. Some authors (e.g., Bartholomew, 1967) contend that the simplicity of analytic models makes them inadequate for studying personnel systems. They maintain that many of the input parameters that are treated as constants by analytic models are not fixed and should be viewed in probabilistic terms using stochastic models. There is no empirical evidence on the model to support or refute this contention.

In conclusion, it appears that the CAPER model will portray realistically a wide variety of personnel systems and will generate valuable information that can be used to aid in personnel planning and decision making.

## REFERENCES

- BARTHOLOMEW, D. J. *Stochastic models for social processes*. New York: Wiley, 1967.
- CRONBACH, L. J., & GLESER, G. C. *Psychological tests and personnel decisions*. (2nd ed.) Urbana: University of Illinois Press, 1965.
- CURTIS, E. W. *The application of decision theory and scaling methods to selection test evaluation*. (Tech. Bull. STB 67-18) San Diego, Calif.: U.S. Naval Personnel Research Activity, February 1967.
- DOPPELT, J. E., & BENNETT, G. K. Reducing the cost of training satisfactory workers by using tests. *Personnel Psychology*, 1953, 6, 1-8.
- DUNNETTE, M. D. Personnel management. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Eds.), *Annual review of psychology*. Palo Alto, Calif.: Annual Reviews, 1962.
- DUNNETTE, M. D. *Personnel selection and placement*. Belmont, Calif.: Wadsworth, 1966.
- MCNEMAR, Q. Moderation of a moderator technique. *Journal of Applied Psychology*, 1969, 53, 69-72.
- SANDS, W. A. Cost of Attaining Personnel Requirements (CAPER) model. Paper presented at the 12th Annual Military Testing Association Meeting, September 14-18, 1970. In H. A. Mahnen & R. C. Willing (Eds.), *Proceedings of the 12th Annual Conference, Military Testing Association*. Indianapolis, Ind.: U.S. Army Enlisted Evaluation Center, 1970.
- SANDS, W. A. *Application of the Cost of Attaining Personnel Requirements (CAPER) model*. (Tech. Bull. WTB 72-1) Washington, D.C.: Naval Personnel Research and Development Laboratory, August 1971. (a)
- SANDS, W. A. *Determination of an optimal recruiting-selection strategy to fill a specified quota of satisfactory personnel*. (Research Memorandum WRM 71-34) Washington, D.C.: Naval Personnel Research and Development Laboratory, April 1971. (b)
- TAYLOR, H. C., & RUSSELL, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 1939, 23, 565-578.
- THORNDIKE, R. L. *Personnel selection: Test and measurement techniques*. New York: Wiley, 1949.
- UHLANER, J. E. *Systems research-opportunity and challenge for the measurement research psychologist*. (Tech. Research Note 108) Washington, D.C.: U.S. Army Personnel Research Office, July 1960.

(Received for Early Publication April 20, 1972)

## RESPONSE REQUIREMENTS AND PRIMACY-RECENCY EFFECTS IN A SIMULATED SELECTION INTERVIEW<sup>1</sup>

JAMES L. FARR<sup>2</sup>

*University of Maryland*

Contrary to findings of Springbett, recency effects of information favorability were found when interviewers made repeated judgments concerning hypothetical applicants for the job of secretary. Consistent order effects were not found when only final judgments were required, although a primacy effect was observed with a rating of overall job suitability in one condition. The obtained recency effects were consistent with data from impression-formation studies. It was suggested that the impression-formation literature might serve as a useful source of selection interview research hypotheses.

One of the earliest findings of the McGill studies of decision making in the selection interview (Webster, 1964) was that information presented early in the interview tended to have greater influence on the final decision than information presented later, that is, a primacy effect (Springbett, 1958). This result was later supported by the research of Sydiah (1961) and Anderson (1960) who investigated interaction processes and speaking times, respectively, in relation to interview decisions.

Data from impression-formation studies employing a similar experimental paradigm are inconsistent with the finding of Springbett. Several studies concerned with methodological and theoretical issues in the information integration aspects of impression formation have found that the number of judgments required of subjects significantly affected the nature of order effects found. Primacy effects were typically found when the subject was required to make only a single judgment after all information had been presented (e.g., Anderson, 1971; Stewart, 1965). Requiring subjects to make repeated judgments based on partial information has generally resulted in recency effects (Byrne, Lambreth, Palmer, & London, 1969; Hendrick & Costantini, 1970; Stewart, 1965).

<sup>1</sup> This article is based on a doctoral dissertation submitted to the Department of Psychology, University of Maryland. Appreciation is expressed to C. J. Bartlett, chairman of the author's doctoral committee and to Frank J. Landy for their careful reading of an earlier version of this article.

<sup>2</sup> Requests for reprints should be sent to James L. Farr, who is now at the Department of Psychology, Pennsylvania State University, University Park, Pennsylvania 16802.

Although the experimental paradigms of Springbett's (1958) study and the various impression-formation investigations were similar, at least two factors distinguish them. First, the type of information presented to the subjects differed. In Springbett's study the information was primarily factual in nature, with the exception of appearance data. In the various impression-formation investigations, the information describing a hypothetical person consisted of a sequence of personality-trait adjectives or a paragraph relating partial daily activities of the person. In general, the impression-formation descriptive information was subjective. Second, the type of judgment required of the subjects differed in the selection interview and impression-formation studies. The subjects in Springbett's (1958) study evaluated job applicants for employment suitability, whereas in the impression-formation studies, the subject typically is asked to judge hypothetical persons with regard to their perceived likableness. The differences in types of information presented and decisions required, or their interaction, may account for the disparate order-effect data.

Webster (1964) hypothesized that the accuracy of interviewer judgments could be improved by presenting the interviewer with a relatively large amount of information concerning an applicant at one time (referred to here as whole presentation) rather than a relatively small amount (sequential presentation). Thus, increasing the amount of information presented at one time might reduce order effects (primacy or recency), as well as other errors of information integration.

The purpose of the present study was to examine the effect of the number of judgments required of subjects, the order of type of information, and the amount of information presented at one time on information favorability order effects. Subjects were also required to make different types of decisions concerning each applicant in order to evaluate the relationship between type of decision and the independent variables.

## METHOD

### Subjects

The subjects were members of the Washington Technical Personnel Forum and were employed in personnel or industrial relations jobs in research and development firms in the Washington, D.C. area. Research materials were mailed to 140 members, and usable replies were received from 77 (55% response rate). In order to equalize sample size for each cell in the experimental design, subjects were randomly withdrawn from those cells with more than 12 responses.

Median characteristics of the subjects were as follows: 37 years old, college graduate, 7 years' experience as an interviewer, and 150 interviews conducted per year.

### Hypothetical Applicants

A total of eight hypothetical applicants were constructed for the job of secretary. Information items used were from the Hakel and Dunnette (1970) list. Items were placed into one of four item pools, using as criteria an item mean favorability rating obtained by Hakel and Dunnette (1970) and item content. An item was classified as high favorability - factual if its mean favorability rating was greater than 5.00 (on a 7-point scale), and the item content was factual. An item was classified as low favorability - factual if it were factual and had received a mean favorability rating less than 3.00. Analogous operations yielded item pools categorized as high favorability - impression and low favorability - impression for those items whose content was impressionistic or yielded personality-trait information. In addition, all items used had been rated as at least moderately important in determining final selection decisions (Hakel & Dunnette, 1970). From the original 730 items of the Hakel and Dunnette list, 38 were classified as high favorability - factual, 53 as low favorability - factual, 64 as high favorability - impression, and 61 as low favorability - impression.

The hypothetical applicants were each constructed of eight items of information. The order of information favorability was varied for each applicant. The orders were designated HHHH, HHLL, HLHL, HLH, LHHL, LHLH, LLHH and LLLL, where H represents two highly favorable items of information and L represents two items of low favorability. Thus, except for applicants HHHH and LLLL, all applicants were constructed of four high- and four low-favorability

items. For all eight applicants four factual and four impressionistic information items were used. Items were nonsystematically selected from the four item pools. The only restrictions on the item sampling procedures were that no item was used for more than one applicant, and no two items for a single applicant could contradict each other (e.g., graduate from college and did not complete high school). The hypothetical applicants were well matched with regard to the favorability of the items from which each was constructed. The average within-applicant item rating for the highly favorable items ranged from 5.59 for applicant HHLL to 5.91 for applicant LLHH. The average within-applicant item rating for items of low favorability ranged from 2.07 for applicant HHLL to 2.23 for applicant LHLH.

### Procedure

Each subject evaluated eight applicants, one of each favorability order. For half of the subjects, the order of type of information presented about the applicants was factual-impressionistic, that is, four factual items followed by four impressionistic ones. The other half were presented the impressionistic-factual order.

The amount of information presented at one time to the subject and the number of judgments required of the subject were treated as a combined variable with three levels. The combined variable was termed presentation-mode-response requirements. The three levels were (a) all information about an applicant presented on a single page with evaluation required after all information was presented (designated whole presentation with final decision); (b) information presented on four pages with two items per page with evaluation required after all information was presented (sequential presentation with final decision); and (c) information presented on four pages with evaluations required after each page (sequential presentation with repeated decisions). One-third of the subjects were placed in each of the three levels.

Three dependent variables were used: ratings of ability to learn the job (termed learning ability), ability to get along with co-workers (termed sociability), and overall suitability for the job of secretary. Each was measured on a 7-point scale (7 = highest evaluation).

The eight hypothetical applicants were placed in a booklet that was mailed to the subjects at their office addresses. The order of applicants was randomized for each subject, and each subject was randomly assigned to experimental conditions.

### Analysis

The original design was a  $2 \times 3 \times 8$  factorial with repeated measurements on the last variable. However, since information favorability order was the primary variable of interest, judgments regarding hypothetical applicants HHHH and LLLL were not used in the primary data analyses. The resulting design was a  $2 \times 3 \times 6$  factorial with repeated measures on the last factor. Separate analyses were conducted for each

TABLE 1  
ANALYSES OF VARIANCE FOR THREE  
DEPENDENT MEASURES

Source of variation	df	MS	F	Omega-squared
Learning ability				
Between				
Presentation-mode-response requirement (A)	2	6.67	3.85*	<.01
Order type information (B)	1	12.34	7.12**	.01
AB	2	.97	.56	.00
Error	66	1.73		
Within				
Order of information favorability (C)	5	73.06	67.85***	.21
AC	10	6.02	5.01*	.03
BC	5	3.43	2.85	<.01
ABC	10	1.99	1.65	<.01
Error	330	1.20		
Sociability				
Between				
A	2	.86	.39	.00
B	1	10.08	4.58*	<.01
AB	2	.72	.33	.00
Error	66	2.20		
Within				
C	5	39.80	40.28***	.11
AC	10	6.06	6.17**	.03
BC	5	29.03	29.54***	.08
ABC	10	2.78	2.83	<.01
Error	330	.98		
Overall suitability				
Between				
A	2	7.11	1.78	<.01
B	1	6.26	1.57	<.01
AB	2	.56	.17	.00
Error	66	3.40		
Within				
C	5	29.80	30.59***	.07
AC	10	4.73	4.85*	.02
BC	5	2.93	3.01	<.01
ABC	10	2.93	3.01	<.01
Error	330	.97		

Note. Within-subject F ratios tested for significance with conservative df (for C, BC: df = 1; 66, for AC, ABC: df = 2; 66).

\* p < .05.

\*\* p < .01.

\*\*\* p < .001.

of the three dependent measures. The within-subject main effect and interaction terms were tested for statistical significance with conservative degrees of freedom as a precaution against heterogeneity of variance and covariance matrices (Myers, 1966, p. 160-162).

The strength of association between the various treatments and the dependent variables was estimated by omega squared (Hays, 1963). Omega squared was calculated by procedures described by Vaughn and Corballis (1969).

#### Definition of Primacy-Recency Effects

For the purposes of this study, a primacy effect of information favorability was said to exist if the hypothetical applicants receiving the highest evaluations were those which were constructed so that highly

favorable information was first presented. Similarly, a recency order effect existed if the applicants rated highest were those constructed so that highly favorable information was presented last.

#### RESULTS

Table 1 presents the analyses of variance for the three dependent measures and the estimates of strength of association between the variables. A consistent finding across dependent variables was that order of information favorability accounted for more of the total variance than any other main effect of interaction. A significant Presentation-Mode-Response Requirements  $\times$  Order of Information Favorability interaction was also found with all three dependent measures.

A significant Order of Information Type  $\times$  Order of Information Favorability interaction was found with only the rating of sociability. The interaction accounted for an estimated 8% of the total variance. Order of information type, although statistically significant for both the rating of learning ability and sociability, accounted for a negligible amount of the total variance. Similarly, a significant main effect of presentation-mode-response requirements accounted for less than 1% of the variance of the rating of learning ability.

Figure 1 presents the data for the Presentation-Mode-Response Requirements  $\times$  Order of Information Favorability interaction for the rating of overall suitability. This figure is illustrative of the cell means obtained for the interaction in question for all dependent measures. A recency effect was found for the sequential presentation with repeated decisions condition for all three ratings. In each instance the three highest rated hypothetical applicants were those with highly favorable information presented last, and of course, the three lowest rated applicants had information of low favorability presented last. No consistent order effects were found in the sequential presentation with final decision condition for any rating measure. A primacy effect was found with the rating of overall suitability in the whole presentation with final decision condition. The highest rated applicants were constructed so that highly favorable information was presented first to the subject. Order effects were not found for the ratings of socia-

bility or learning ability in the whole presentation condition.

An examination of the Order of Information Type  $\times$  Order of Information Favorability interaction revealed that impressionistic information was more important than factual information in determining the evaluation of sociability. When the impressionistic information was of low favorability, the applicant was rated low in sociability. The opposite evaluation occurred when the impressionistic information was highly favorable. Thus, information content may interact with favorability in affecting some judgment decisions.

### DISCUSSION

The data of the present study generally supported the findings of impression-formation studies that used similar experimental paradigms. A recency effect of information favorability was found for all dependent variables when repeated judgments were required of the subjects. Analogous findings have been reported by several researchers in impression formation (Byrne et al., 1969; Hendrick & Costantini, 1970; Stewart, 1965). The findings of Springbett (1958) were not supported by these data. At least two factors distinguish the present study from that of Springbett. In Springbett's research, judgments were required only after a fairly large amount of information about an applicant had been presented, whereas in the present case only eight items of information were presented about each applicant. It appears reasonable that the amount of information presented before a preliminary judgment is required may moderate whether primacy or recency effects are found.

Springbett (1958) used a dichotomous rating scale with categories of "accept" and "reject." The present study, however, used a 7-point response scale, allowing more gradations of judgment. The ranking of applicants, as measured on a multichotomous scale, could change substantially without necessarily affecting judgments of accept versus reject. Thus, recency effects as defined in the present research possibly could have been exhibited in Springbett's data but were masked by the response dichotomy.

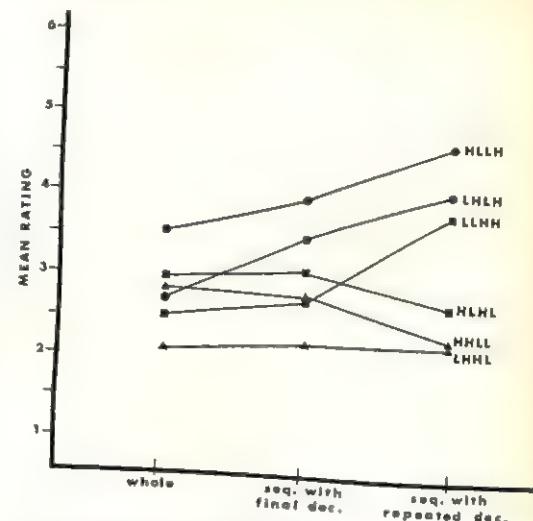


FIG. 1. Presentation-Mode-Response Requirements  $\times$  Order of Information Favorability interaction: Rating of overall sociability.

Anderson (1971) explains primacy and recency effects found in various experimental conditions in impression-formation studies by an attention hypothesis. When only a final judgment is required, primacy effects result from the decreased attention paid to information presented later to the evaluator. The attention hypothesis explains recency effects when repeated judgments are required by proposing that the additional response requirements force an increase in attention to the later information.

The tenability of the attention hypothesis would have practical implications in the interview context. The utilization of repeated judgments about an applicant would increase the likelihood that the interviewer attended to information presented relatively late in the interview. This increased attention could be detrimental or facilitative with regard to accuracy of prediction. To maximize the usefulness of the interview would require the pairing of the judgment with information of high importance and validity. To match the later judgments with information of low importance or validity would lead to a decrease in judgment accuracy. Thus, in a practical interview situation, the validity of various types of information would have to be known in order to use satisfactorily repeated judgments as a means of increasing attention to all information about a job applicant.

The finding that certain types of information are more important than others in determining specific evaluation responses may also be of practical significance. If an interview is used for a limited purpose, such as predicting sociability of the applicant (as suggested by Ulrich & Trumbo, 1965), then the information given to the interviewer should be restricted to that which is important to the decision at hand. To present irrelevant information would only tend to lower the quality of the final decision reached.

The estimated proportion of total variance accounted for by the various experimental conditions in the present study was lower than that reported in other investigations (e.g., Carlson, 1971; Hakel, Dobmeyer, & Dunnette, 1970). Table 1 indicates that only 25%, 22% and 9% of the variance of the ratings of learning ability, sociability, and overall suitability, respectively, were accounted for. The exclusion from the data analyses of the hypothetical applicant constructed of all favorable and the one constructed of all unfavorable information lowered the estimates of accountable variance. Analyses of variance conducted with responses to all eight applicants indicated that 37%, 37%, and 33% of the total variance of the ratings of learning ability, sociability, and overall suitability, respectively, were accounted for. These proportions were more comparable to those of the previously mentioned studies.

#### REFERENCES

ANDERSON, C. W. The relation between speaking times and decision in the employment interview. *Journal of Applied Psychology*, 1960, 44, 267-268.

- ANDERSON, N. H. Integration theory and attitude change. *Psychological Review*, 1971, 78, 171-206.
- BYRNE, D., LAMBRETH, J., PALMER, J., & LONDON, O. Sequential effects as a function of explicit and implicit interpolated attraction responses. *Journal of Personality and Social Psychology*, 1969, 13, 70-78.
- CARLSON, R. E. Effect of interview information in altering valid impressions. *Journal of Applied Psychology*, 1971, 55, 66-72.
- HAKEL, M. D., DOBMEYER, T. W., & DUNNETTE, M. D. Relative importance of three content dimensions in overall suitability ratings of job applicants' resumés. *Journal of Applied Psychology*, 1970, 54, 65-71.
- HAKEL, M. D., & DUNNETTE, M. D. *Checklists for describing job applicants*. Minneapolis: University of Minnesota Industrial Relations Center, 1970.
- HAYS, W. L. *Statistics for psychologists*. New York: Holt, Rinehart & Winston, 1963.
- HENDRICK, C., & COSTANTINI, A. F. Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. *Journal of Personality and Social Psychology*, 1970, 15, 158-164.
- MYERS, J. L. *Fundamentals of experimental design*. Boston: Allyn & Bacon, 1966.
- SPRINGBETT, B. M. Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology*, 1958, 12, 13-22.
- STEWART, R. H. Effect of continuous responding on the order effect in personality impression formation. *Journal of Personality and Social Psychology*, 1965, 1, 161-165.
- SYDIAHA, D. Bales' interaction process analysis of personnel selection interviews. *Journal of Applied Psychology*, 1961, 45, 393-401.
- ULRICH, L., & TRUMBO, D. The selection interview since 1949. *Psychological Bulletin*, 1965, 63, 100-116.
- VAUGHN, G. M., & CORBALLIS, M. C. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 1969, 72, 204-213.
- WEBSTER, E. C. *Decision-making in the employment interview*. Montreal: Industrial Relations Center, McGill University, 1964.

(Received December 20, 1971)

# TRAINING INTERVIEWERS TO ELIMINATE CONTRAST EFFECTS IN EMPLOYMENT INTERVIEWS

KENNETH N. WEXLEY,<sup>1</sup> RAYMOND E. SANDERS, AND GARY A. YUKL

*University of Akron*

Contrast effects have been found to be a potentially serious source of error in interviewers' ratings of job applicants. A series of experiments was conducted in an attempt to eliminate these errors. Use of a warning in the first experiment was not successful. Use of an anchoring treatment in the second experiment was equally unsuccessful. Combining and strengthening the warning and anchor treatments in the third experiment also failed, revealing that contrast effects are a surprisingly tenacious source of rating error. Finally, in the fourth experiment, an intensive workshop incorporating basic learning principles was successful in eliminating contrast effects as well as some other sources of interviewer error.

The influence of contrast effects on employment interviewers' ratings of job applicants has been found in several recent studies (Carlson, 1970; Hakel, Ohnesorge, & Dunnette, 1970; Leonard & Hakel, 1971). These studies have demonstrated that interviewers' evaluations of job applicants can be affected by the suitability of immediately preceding applicants. Wexley, Yukl, Kovacs, and Sanders (1972) found that the magnitude of these contrast effects is greatest with applicants of intermediate suitability and could account for as much as 80% of the variance in ratings. Therefore, it is important to find ways to reduce or eliminate this potential source of rating error. This article describes a series of experiments in which an attempt was made to substantially reduce contrast effects in interviewer ratings by (*a*) warning interviewers about this source of error and (*b*) providing absolute standards as anchors.

## EXPERIMENT I

The purpose of the first experiment was to determine whether the influence of contrast effects, as a source of error in employment interview ratings of job applicants, can be substantially reduced by warning interviewers about them.

<sup>1</sup> The authors would like to thank Peter J. Hunt for his help in collecting data for Experiments I, II, and III.

Requests for reprints should be sent to Kenneth N. Wexley, Department of Psychology, University of Akron, Akron, Ohio 44304.

## Method

The subjects watched videotaped interviews of hypothetical applicants for a sales job and rated these applicants in terms of their qualifications on a 9-point rating scale. The development of the videotapes has been discussed elsewhere (Wexley et al., 1972).

Each subject was shown three videotapes. The first two videotapes were used to establish either a high (H) or low (L) frame of reference in the subjects. The third videotape always showed an average (A) suitability applicant. Thus, two experimental conditions (i.e., HHA and LLA) were used. Ratings of the average applicant were analyzed to determine the amount of contrast effects. The subjects were 20 undergraduate psychology students who were paid for their participation. Ten subjects were randomly assigned to each experimental condition.

Before seeing the first applicant, subjects were given a detailed description of the sales job, a list of the qualifications needed for the job, and an evaluation guide. The evaluation guide consisted of a list of questions about an applicant's qualifications which the subjects were asked to consider before rating him.

The procedure up to this point will be referred to as the "test phase." The test phase was used in all four experiments, each experiment with 20 new subjects.

Prior to seeing the first videotape, the subjects were given the following warning: "It has been found in earlier research that the evaluation of a particular applicant's job suitability is influenced by the job suitability of previously interviewed applicants. Therefore, please make sure that you rate each applicant on his own merit and not on how he compares to those applicants interviewed before him."

## Results and Discussion

It is immediately apparent from the results in Table 1 that the warning, despite its potential demand effect (Orne, 1963), had

TABLE 1  
MEANS, STANDARD DEVIATIONS, AND VARIANCE ANALYSES  
FOR THE RATINGS IN EACH EXPERIMENT

Condition	<i>M</i> (LLA)	<i>M</i> (HHA)	<i>SD</i> (LLA)	<i>SD</i> (HHA)	<i>MS</i> between	<i>MS</i> within	<i>F</i>	% Variance
Wexley et al. (1970)	8.1	2.5	0.70	1.86	156.80	2.19	71.60*	80%
Warning (Experiment I)	7.1	2.4	1.64	1.95	115.20	3.61	31.91*	64%
Anchoring (Experiment II)	6.4	2.7	1.74	1.10	68.45	2.36	29.00*	62%
Combination (Experiment III)	7.0	3.2	1.67	1.40	72.20	2.09	34.55*	66%
Workshop (Experiment IV)	5.2	4.8	1.66	0.60	0.80	1.73	0.46	3%

\*  $p < .01$ .

little impact on reducing contrast effects in the subjects' ratings. The contrast effects were statistically significant, and they continued to account for a substantial part (64%) of the total variance in the ratings. Thus, it appears that even when subjects are warned to avoid contrast effects, their judgments still fall victim to this source of error.

## EXPERIMENT II

The purpose of Experiment II was to reduce contrast effects by providing subjects with some absolute standards. This was attempted by anchoring subjects at the two extreme ends of the 9-point rating scale.

### Method

In the second experiment, the subjects were given an anchoring treatment instead of a warning prior to seeing the first applicant. Two anchor stimuli were used, one representing a high suitability applicant and one representing a low suitability applicant. Each anchor stimulus consisted of a written summary of an applicant's responses from one of the extra videotapes used in the earlier study (Wexley et al., 1972).<sup>2</sup>

The subjects read the two anchor descriptions to themselves while the experimenter read them aloud. The subjects were told that the high suitability anchor represented a "9" on the rating scale and the low suitability anchor represented a "1" on the rating scale. Following the anchoring procedure, they were administered the test phase.

### Results and Discussion

Examination of the ANOVA data for Experiment II gives essentially the same results

as in Experiment I (see Table 1). Anchoring of subjects was not very effective in reducing contrast effects, which persisted in accounting for a sizeable 62% of the total variance in the subjects' evaluations.

## EXPERIMENT III

Since the first two experiments were unsuccessful in reducing contrast effects to any large extent, in the third experiment the warning treatment was combined with the anchoring treatment, and both treatments were modified in an attempt to strengthen them.

### Method

The method was generally the same as in the previous experiments. However, the anchoring treatment was strengthened by including an average suitability applicant as an additional anchor stimulus. Moreover, the experimenter pointed out to the subjects the strengths and weaknesses of each anchor applicant. This was done by reviewing the list of questions in the evaluation guide regarding each applicant's qualifications. The warning treatment was strengthened by giving the subjects examples of how leniency, halo, central tendency, contrast, and stereotyping can distort a judge's ratings in a beauty contest. At this point, the test phase was administered to the subjects.

### Results and Discussion

Table 1 reveals that the proportion of rating variance due to contrast effects in Experiment III was 66%. It is obvious from these results that combining and strengthening the treatments of Experiments I and II was still unsuccessful in reducing contrast errors to any appreciable degree.

<sup>2</sup> A pilot study revealed that videotapes and written summaries have identical effects as anchor stimuli; we decided to use the written summaries because they required less time to administer.

#### EXPERIMENT IV

The first three experiments indicated that contrast effects are surprisingly difficult to reduce to any substantial degree. Despite all initial efforts, contrast errors persisted in accounting for 62-66% of the decision variance. In seeking a procedure for training interviewers to be less susceptible to this rating error, a 2-hour workshop was developed by the authors. The objective of Experiment IV was to determine whether this workshop could be effective in reducing the magnitude of contrast effects.

##### *Method*

Four separate workshop sessions were held, each with five subjects. They were introduced to one another and to the two experimenters who acted as trainers. The subjects were told that they were to participate in a workshop designed to improve their skill as employment interviewers. They were also informed that the purpose of the study was to evaluate the effectiveness of the workshop, not to test them in any way.

A job description and a list of the applicant qualifications required for the job were given to each subject. After reading these handouts, the subjects were asked to put them face down and, as a group, to discuss the duties of the job, the qualification(s) needed to perform each job duty successfully, and the way one can use an interview to determine whether an applicant meets each qualification. The subjects were then given the evaluation guide which they were permitted to refer to while they watched and rated the applicants.

Each group of subjects was asked to watch three videotaped employment interviews.<sup>3</sup> One tape showed an H applicant, one showed an L applicant, and one showed an applicant of A suitability. Two groups saw these three training videotapes in a LAH sequence while two groups saw them in an HAL sequence. After watching an applicant, the subjects individually rated him on the 9-point rating scale and then announced their rating to the group. In addition, each subject explained to the group his reasons for giving that particular rating. The subjects also discussed possible reasons for the discrepancies among their ratings. During these discussions, the trainers announced what the correct rating should have been for that particular applicant (i.e., either a rating of 9, 5, or 1). Moreover, experimenters made reference to various types of rating errors including leniency, halo, central tendency, contrast, and stereo-

<sup>3</sup> The videotapes were used instead of written summaries because we believed that a more realistic presentation of the practice applicants would facilitate training; the three videotapes were from the earlier study by Wexley et al.

typing. The experimenters pointed out a given type of rating error at the time when one or more subjects actually committed it.

The workshop lasted about 2 hours, after which there was a 10-minute break. Following the break, subjects were administered the test phase.

##### *Results*

From the results (see Table 1), it is clear that the workshop was successful. Contrast errors were not statistically significant and accounted for only 3% of the decision variance. In addition, the pattern of mean squares in Table 1 indicates that the workshop succeeded in reducing the MS within groups as well as the MS between groups. The reduction in MS within groups was probably due to the successful reduction of other types of rating errors besides contrast effects. Careful questioning of the subjects following each training session showed that they were unaware of the actual purpose of the workshop. Thus, the workshop's success did not appear to be contaminated by demand effect.

##### *DISCUSSION*

The results of Experiments I, II, and III show the tenacity of contrast effects despite attempts to reduce them by means of warning and/or anchoring of subjects. The results of Experiment IV indicate that contrast effects can only be reduced by a fairly intensive training program which takes into account some of the basic principles of learning. Our training workshop gave subjects a chance to practice observing and rating actual videotaped applicants, provided subjects with immediate feedback concerning the accuracy of their ratings, and maintained the subjects' interest by using realistic stimuli and by encouraging informal group discussions. Training workshops of this type appear to be a practical and relatively inexpensive method for training interviewers in industry.

Two possible limitations of these experiments should be mentioned. First, it is not clear exactly what features of the workshop were responsible for its success. Second, the subjects used were students, not actual employment interviewers. Whether workshops of this nature will be as effective with experienced interviewers must await further research.

## REFERENCES

- CARLSON, R. E. Effect of applicant sample on ratings of valid information in an employment setting. *Journal of Applied Psychology*, 1970, 54, 217-222.
- HAKEL, M. D., OHNESORGE, J. P., & DUNNETTE, M. D. Interviewer evaluations of job applicants' resumés as a function of the qualifications of the immediately preceding applicants: An examination of contrast effects. *Journal of Applied Psychology*, 1970, 54, 27-30.
- LEONARD, R. L., JR., & HAKEL, M. D. Contrast and assimilation effects in an employment setting: A theoretical re-interpretation. Paper presented at Midwestern Psychological Association Convention, Detroit, May 1971.
- ORNE, M. T. On the social psychology of the psychology experiment; with particular reference to demand characteristics and their implications. *American Psychologist*, 1963, 17, 776-783.
- WEXLEY, K. N., YUKL, G. A., KOVACS, S. Z., & SANDERS, R. E. Importance of contrast effects in employment interviews. *Journal of Applied Psychology*, 1972, 56, 45-48.

(Received November 17, 1971)

## EFFECT OF RACE ON PEER RATINGS IN AN INDUSTRIAL SITUATION<sup>1</sup>

FRANK L. SCHMIDT<sup>2</sup> AND RAYMOND H. JOHNSON

*Michigan State University*

The effect of race on peer ratings was examined in an industrial sample which was approximately 50% black and which had recently been exposed to training in human relations. Contrary to results in previous studies, no race effect was found. In addition, almost all the requirements for convergent and discriminant validity between the races were met. Possible explanations for these results and implications for the use of peer ratings in integrated settings were discussed.

A number of studies have aimed at clarifying the effects of certain characteristics of raters, ratees, and situations on peer ratings and nominations. For example, it has been found that, while friends appear to be favored with higher peer nominations, the validity of such nominations is not adversely affected (Hollander, 1956; Waters & Waters, 1970; Wherry & Fryer, 1949), and partialing the effect of friendship out of the nominations seems to leave validities virtually unchanged. Lewin, Dubno, and Akula (1971) found face-to-face interaction was apparently not critical in peer ratings; ratings made after watching ratees on videotape were almost identical to those made after fairly extensive face-to-face interaction. Length of acquaintance in face-to-face situations does not appear to affect reliability of peer ratings; reliabilities of ratings made after 3-4 days acquaintance were similar in size (.80s and .90s) to those of ratings made after longer acquaintance (Hollander, 1957); and peer ratings made of individuals by the same group of peers seem to be stable over periods of up to 2 years (Wodder, 1962). It has even been found that peer ratings given to an individual are stable when the individual is moved from group to group within an orga-

nization (Gordon & Medland, 1965; Medland & Olans, 1964). Each of these studies underlines the relative lack of effect of acquaintanceship factors on the reliability and validity of peer ratings.

With the increasing incidence of racially integrated industrial work groups, it becomes important to know if such findings extend to the variable of race. Only two previous studies examining the effect of race on peer ratings were found. Both Cox and Krumboltz (1958) and DeJung and Kaplan (1962) found that raters gave significantly higher ratings to ratees of their own race than to those of the other race and that this effect was more marked for the black than white raters.<sup>3</sup> Nevertheless, black and white raters showed fairly high intercorrelations. In the Cox and Krumboltz study, the correlations between ratings produced by the two races for ratings of leadership ability given blacks, whites, and the total group were .75, .77, and .75 respectively; DeJung and Kaplan (1962) found between-race Interrater correlations for ratings of combat potential to be .42 and .47 for two black groups and .52 for each of the two white groups.

These two studies share a number of characteristics. Both were carried out in military settings, the earlier study in the Air Force and the more recent one in the Army. In both studies, blacks constituted only a relatively small percentage of the peer groups studied,

<sup>1</sup> This study was supported by the Chrysler Institute, Chrysler Corporation, Detroit, Michigan. The researchers are especially grateful to Dennis J. Deshaies for his support and assistance. They would also like to acknowledge the contributions of D. L. Maxwell, W. R. DeBusk, C. V. Roman, D. J. Lewisley, J. G. Hafner, A. M. Gray, and J. M. Hall.

<sup>2</sup> Requests for reprints should be sent to Frank L. Schmidt, Department of Psychology, Michigan State University, East Lansing, Michigan 48823.

<sup>3</sup> Flaugh, Campbell, and Pike (1969) have found a similar race effect in ratings made by supervisors, but the psychological processes involved may be quite different from those involved in the race effect in peer ratings.

possibly resulting in a situation in which a black rating other blacks was usually rating his closest friends. The white, on the other hand, in rating members of his own racial group was rating nearly all of his peers, diluting the effect that would result if higher ratings were given to close friends. If such a friendship effect were operating, and if friendships tended to be race bounded, the greater race effect shown by black raters could have been traceable to the numerical imbalance between the races in the peer groups. Finally, both studies included peer ratings on only one trait, and thus did not allow assessment of discriminant validities.

The present study was an attempt to ascertain whether the race effect would be found in an industrial training setting in which blacks constituted roughly 50% of the peer groups and in which raters had recently been exposed to training emphasizing interracial fairness and understanding. The inclusion of two traits allowed the assessment of both convergent and discriminant validity of peer ratings from raters of different races via the multitrait-multimethod matrix of Campbell and Fiske (1959). This method has proven useful in prior studies in assessing the general construct validity of ratings produced by different categories of raters (Gunderson & Nelson, 1966; Lawler, 1967; Thompson, 1970).

#### METHOD

Subjects were 43 black and 50 white trainees in an experimental foreman-training program in a large midwestern manufacturing concern. Selection for the program was exclusively from the ranks of present hourly employees and was based on self-nomination, ability test scores, superiors' recommendations, and past work record. Average educational level was slightly above 12 years for both races. Mean age was 29.06 for whites and 31.09 for blacks. Subjects underwent training in six groups ranging in size from 11 to 24. In addition to a week of traditional lecture-oriented training, the men received 40 hours of intensive human relations training. The techniques of sensitivity training were combined with role playing exercises, immediate video feedback, and eclectic discussions of human relations principles. Racial differences and conflicts were aired and discussed whenever they arose.

As part of a larger study to evaluate this program, the men in each training group were asked to rate their fellow trainees on two traits, using a five-

category forced-distribution rating scale. After crossing his own name off the list of group members, each trainee distributed his peers into the top 10%, next 20%, middle 40%, next 20%, and lowest 10% on each of two traits: (a) predicted future success as a foreman and (b) general drive and assertiveness. Descriptions of these traits are given in the following excerpts from the instructions read to the subjects.

*Drive and assertiveness.* One trait we would like you, as a trainee, to rate your fellow trainees on is general assertiveness, pushing of self, or drive. A person high on this trait appears to be energetic, motivated, and self-confident. He takes the lead in discussions and in organizing tasks. People low in this trait, on the other hand, are somewhat shy and lacking in self confidence. They tend to be less aggressive and to speak up less often in group discussions.

*Future success.* We would like you to estimate how successful your fellow trainees will be later on as foremen, when they actually have to deal with the day-to-day problems of a first-line supervisor. Do not base evaluations on how well the person has done in training but instead on how well you think he will actually do as a foreman later when he is on the job.

Trainees in each group were instructed as to the number of names that had to be placed in each of the five categories of the scale and were assured that their ratings were to be used for research purposes only and would not in any respect affect their futures with the company or the futures of their peers.

#### Analysis

By treating each rater as an "item," reliabilities<sup>4</sup> were computed separately in each of the six training groups for each trait. These ratings were of (a) the whole group by blacks, (b) the whole group by whites, (c) blacks by blacks, (d) blacks by whites, (e) whites by blacks, and (f) whites by whites.

Ratings given by blacks and by whites were based on an average across groups of 6.64 and 8.36 raters, respectively. In order to allow comparison of reliabilities between the races, the average of these two figures (7.50) was used in the Spearman-Brown formula to adjust each of the 12 coefficients in each of the six groups. Reliabilities were then averaged across groups to obtain final estimates.

For each trait a two-factor analysis of variance was employed to assess the effect of race. There were two levels of each factor: black versus white raters and black versus white ratees, with repeated measures on the ratee factor. Three blacks and 10 whites were discarded randomly for this analysis to provide equal *Ns* of 40 blacks and 40 whites. *F*<sub>MAX</sub> tests

<sup>4</sup> Internal consistency reliabilities—since "item" responses were continuous, coefficient alpha rather than Kuder-Richardson formula 20 was the appropriate form.

indicated that the assumption of homogeneity of variance was met.

Three separate multitrait-multimethod matrices were constructed. The first contained ratings given to the combined group; the second, ratings given to blacks; and the third, ratings given to whites. This breakdown allowed for examination of differences in convergent and discriminant validity as a function of the ratee sample. It was expected that the two traits rated would show a moderately high positive correlation under all conditions and that, for this reason, the requirements of discriminant validity would be somewhat more difficult to meet than is usually the case (Campbell & Fiske, 1959, p. 103).

### RESULTS AND DISCUSSION

Table 1 presents the means and standard deviations of ratings on both traits assigned by each race to both races. The kind of race effect found by Cox and Krumboltz (1958) and DeJung and Kaplan (1962) would require that raters rate members of their own race higher than members of the other race (i.e., that there be a significant ratee by rater interaction) and that this effect be more marked for black than white raters (which would result in a significant ratee effect). In Table 1, it can be seen that there is a tendency for raters to rate same-race ratees higher in predicted future process than different-race ratees, but this effect is greater for white than black raters. Both white and black raters gave slightly higher mean ratings on drive and assertiveness to blacks; however, neither the interaction nor the ratee main effect reached significance in either of the analyses of variance.

Of the three factors in this study which differed from those in past studies, it seems unlikely that the fact of a civilian rather than military setting would produce a strong effect in the direction of eliminating the race effect. If DeJung and Kaplan's (1962) hypothesis concerning race-bound friendships has validity, the critical variable accounting for the absence of a race effect may be the relatively large proportion (46.2%) of blacks in these peer groups, which created approximately equal probabilities that black and white raters rating same-race ratees are rating their close friends. The design of the study does not allow for separation of the effect of the friendship variable from the effect, if any, of the human relations training.

TABLE 1  
MEAN RATINGS AND STANDARD DEVIATIONS OF RATINGS ASSIGNED BY RATERS OF BOTH RACES TO RATEES OF BOTH RACES

Trait	White rater	Black rater
Predicted future success		
Black		
$M$	2.98	3.03
$SD$	.55	.65
White		
$M$	3.10	2.97
$SD$	.58	.54
Drive and assertiveness		
Black		
$M$	3.09	3.10
$SD$	.61	.64
White		
$M$	3.00	2.94
$SD$	.56	.66

Tables 2, 3, and 4 present the multitrait-multimethod matrices for the ratee group as a whole, for black ratees, and for white ratees, respectively. The monotrait-heteromethod correlations are significant and large in all three matrices, thus meeting the requirement for convergent validity. The discriminant validity requirement that each convergent validity be higher than the values lying in its column and row in the heterotrait-heteromethod matrix is met by all convergent validities in the three matrices. A second requirement for discriminant validity is that the convergent validity coefficient for each variable should be larger than the correlation between this variable and other variables in the heterotrait-monomethod triangles. Because of the pervasiveness of method variance, this requirement is seldom met by behavioral data (Gunderson & Nelson, 1966; Lawler, 1967; Thompson, 1970), even though it is usually interpreted to mean only that the *average* of the heterotrait-monomethod correlations must be smaller than the average of the convergent validity coefficients. In this data, the relatively high heterotrait-monomethod correlation produced by the white raters in each of the three matrices precludes satisfying the more stringent of the two conditions, although, in each case, even this requirement is almost met. In all three matrices the *average* of the convergent validity coeffi-

TABLE 2

## MULTITRAIT-MULTIMETHOD MATRIX FOR BLACK AND WHITE RATERS WHEN RATING COMBINED SAMPLE

Trait	1	2	3	4
Method 1 (black raters)				
Predicted future success (1)	(.70)			
Drive and assertiveness (2)	.61	(.85)		
Method 2 (white raters)				
Predicted future success (3)	.70	.64	(.83)	
Drive and assertiveness (4)	.52	.77	.71	(.82)

Note. The monotrait-heteromethod correlations are in italics.

clients exceeds the average of the heterotrait-monomehtod correlation (.74 vs. .66 in Table 2; .72 vs. .58 in Table 3; and .76 vs. .75 in Table 4), but this less stringent requirement is only marginally met in the ratings given to whites. In view of the fact that predicted future success and drive and assertiveness were considered to be related concepts and were expected to show a relatively high intercorrelation and the fact that no studies with behavioral data could be found in which even this relaxed criterion was satisfied, the extent to which the present data meet this requirement appears quite adequate.

A third condition for discriminant validity is that the same pattern of correlations appear in all of the heterotrait triangles of both the monomehtod and the heteromethod blocks. Since a minimum of three traits is necessary to assess these patterns, this requirement cannot be applied to these data. A final condition, that the reliability of each variable

be higher than its heterotrait-monomehtod correlations, is, with one exception, met for both traits in all three matrices.

In Table 4 it can be seen that the ratings by blacks of whites on predicted future success show a reliability smaller than the intertrait correlation in that monomehtod block. This indicates perhaps that the black raters did not perceive predicted future success and drive and assertiveness as two separate traits in the white ratees. In black ratees, on the other hand, black raters seemed to see these traits as less related than did white raters (see Table 3).

Extent of method variance is indicated by the difference in level of correlation between the parallel values of the monomehtod block and the heteromethod block (Campbell & Fiske, 1959). According to this yardstick, very little method variance due to race is evident in these data.

In general, these data meet the requirements for convergent and discriminant validity quite well. The peer ratings made by the two races in this study can quite safely be considered comparable methods of assessing these two traits.

In summary, these findings indicate that the racial bias effect in peers ratings does not inevitably occur and that an approximately equal proportion of minority and majority group members in peer groups and/or human relations training may be associated with its nonoccurrence. In addition, black and white raters were found to show relatively high levels of discriminant and convergent validity in assessing black ratees, white ratees, and

TABLE 3

## MULTITRAIT-MULTIMETHOD MATRIX FOR BLACK AND WHITE RATERS WHEN RATING BLACKS ONLY

Trait	1	2	3	4
Method 1 (black raters)				
Predicted future success (1)	(.83)			
Drive and assertiveness (2)	.48	(.72)		
Method 2 (white raters)				
Predicted future success (3)	.65	.55	(.80)	
Drive and assertiveness (4)	.44	.78	.67	(.82)

Note. The monotrait-heteromethod correlations are in italics.

TABLE 4

## MULTITRAIT-MULTIMETHOD MATRIX FOR BLACK AND WHITE RATERS WHEN RATING WHITES ONLY

Trait	1	2	3	4
Method 1 (black raters)				
Predicted future success (1)	(.66)			
Drive and assertiveness (2)	.73	(.74)		
Method 2 (white raters)				
Predicted future success (3)	.76	.73	(.86)	
Drive and assertiveness (4)	.60	.76	.77	(.81)

Note. The monotrait-heteromethod correlations are in italics.

combined groups. The implication is that the highly valid prediction device of peer ratings may be quite appropriate and useful in many integrated situations. Future research might well focus on the relative potency of training in human relations, the proportion in the peer group that is minority, and other factors in contributing to the elimination of the racial bias effect in peer ratings.

## REFERENCES

- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.
- COX, J. A., & KRUMBOLTZ, J. D. Racial bias in peer ratings of basic airman. *Sociometry*, 1958, 21, 292-299.
- DEJUNG, J. E., & KAPLAN, H. Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. *Journal of Applied Psychology*, 1962, 46, 370-374.
- FLAUGHER, R. L., CAMPBELL, J. T., & PIKE, L. W. Prediction of job performance for Negro and white medical technicians: Ethnic group memberships as a moderator of supervisors ratings. (ETS Service Report PR-69-5) Princeton: Educational Testing Service, 1969.
- GORDON, L. V., & MEDLAND, F. I. The cross-group stability of peer ratings of leadership potential. *Personnel Psychology*, 1965, 18, 173-177.
- GUNDERSON, E. K. E., & NELSON, P. D. Criterion measures for extremely isolated groups. *Personnel Psychology*, 1966, 19, 67-80.
- HOLLANDER, E. P. The friendship factor in peer nominations. *Personnel Psychology*, 1956, 9, 435-447.
- HOLLANDER, E. P. The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology*, 1957, 41, 85-90.
- LAWLER, E. Multitrait-multirater approach to measurement of job performance. *Journal of Applied Psychology*, 1967, 51, 369-381.
- LEWIN, A. Y., DUBNO, P., & AKULA, W. G. Face-to-face interaction in the peer-nomination process. *Journal of Applied Psychology*, 1971, 55, 495-497.
- MEDLAND, F. F., & OLANS, J. L. Peer rating stability in changing groups. (USA PRO Tech. Res. Note No. 142) Washington, D.C.: U.S. Army Personnel Research Office, 1964.
- THOMPSON, H. A. Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach. *Journal of Applied Psychology*, 1970, 54, 496-502.
- WATERS, L. K., & WATERS, C. W. Peer nominations as predictors of short-term roles performance. *Journal of Applied Psychology*, 1970, 54, 42-44.
- WILERRY, R. J., & FRYER, D. H. Buddy ratings: Popularity contest or leadership criterion? *Personnel Psychology*, 1949, 2, 147-159.
- WODDER, N. C., & HALL, W. E. An analysis of peer ratings. *Personnel Guidance Journal*, 1962, 40, 606-609.

(Received November 15, 1971)

## AN APPROACH FOR DETERMINING CRITERIA OF SALES PERFORMANCE

DAVID W. CRAVENS<sup>1</sup> AND ROBERT B. WOODRUFF

*College of Business Administration, University of Tennessee, Knoxville*

The focus of this study was on the design and testing of a methodology for analytically determining standards of sales performance. The methodology consists of (a) formulating a conceptual model of sales territory performance, (b) selecting variables and corresponding operational measures for a given organization, and (c) empirically determining the degree to which these measures explain variations in territory performance. The salesman performance as assessed by the firm's management appeared to be consistent with the analytically determined performance standards.

While the need for methods of predicting and evaluating the performance of salesmen is great, previous research has not been particularly successful in identifying variables associated with salesman performance (Baehr & Williams, 1968; Cotham, 1968; Miner, 1962).

Researchers have focused on a wide variety of predictor variables thought to explain differences in the performance of salesmen in sales organizations. Yet, far less attention has been given to determining appropriate performance criterion variables (Cotham, 1970). The unimpressive results of previous research in this area may be due, at least in part, to insensitive measures of salesman performance.

Since the salesman is only one of many factors influencing sales territory results, criterion variables such as sales volume or sales-based ratios measure sales territory performance rather than salesman performance unless standards are adjusted for factors beyond the salesman's control. (A sales territory is the responsibility assigned to a single salesman in terms of products, customers, store department, or geographical area.) Salesman-oriented predictor variables such as personal history items, personality traits, and attitudes should not be expected to correlate highly with criterion variables that are partially influenced by factors not under the personal control of the salesman.

An approach for determining measures of salesman performance consists of identifying

<sup>1</sup> Requests for reprints should be sent to David W. Cravens, College of Business Administration, University of Tennessee, Knoxville, Tennessee 37916

determinants of territory performance in a given organization, selecting operational measures of these determinants, and then empirically determining the degree to which these measures explain variation in territory performance. If most of the variation can be explained by this procedure, then by separately analyzing only those factors beyond the control of the salesman, a performance standard or benchmark for each territory can be generated. Comparisons of actual results against these standards can be used as criterion measures in future research attempting to identify predictors of salesman performance.

The research presented in this article seeks to provide greater insight into the determination of valid performance measures. A conceptual model is presented to show the variety of factors affecting performance in sales territories. An exploratory study was then conducted to separate the role of salesmen from that played by other factors in the sales territories of a large durable goods manufacturer. The resulting operational implications for predicting and evaluating salesman performance are discussed.

### METHOD

#### *Research Site*

The field sales organization used in the study was an international manufacturer of high priced consumer goods. Twenty-five territories and the salesmen assigned to them, comprising approximately 30% of the entire sales organization, were included in the analysis. They represented a distinct geographical area and ranged from smaller, more congested territories to larger territories such that a representative

cross section of areas was obtained. Moreover, more complete historical information was available for the territories comprising this particular sector of the firm's nationwide organization.

### *Conceptual Model*

The many factors other than the salesman which can influence results in a sales territory are well known to experienced sales management. Nevertheless, it is helpful to combine the factors into a conceptual model as an aid to selecting appropriate variables for study in a particular organization. Thus, the following conceptualization is simply a composite of existing knowledge in the field.

At least three general influences can affect territory performance: (a) factors which have the same impact on all territories (e.g., a nationwide strike in a firm's production plants), (b) factors which affect only one or a few territories (e.g., a disastrous hurricane in a coastal area), or (c) factors which affect all territories but not to the same degree (e.g., the degree of market opportunity present or the experience of assigned salesmen).

Influences falling into the first category should have no effect on interterritory comparisons assuming their impact is constant throughout the sales organization. The second category can be handled by eliminating territories affected or by interpreting their performance in terms of the situation-specific influences. The third category of factors is likely to account for major sources of variation between territories and thus provides the focus for analysis. These influences can be expressed using the following composite categories:

$$TP = f(P, W, S_r, S_f, C_r, C_f, O)$$

where:  $TP$  = territory performance,  $P$  = industry market potential (in territory),  $W$  = territory workload,  $S_r$  = salesman experience,  $S_f$  = salesman motivation and effort,  $C_r$  = company experience in territory,  $C_f$  = company effort in territory, and  $O$  = other factors.

This relationship is stated in a general form since determination of the specific relationship that best describes a particular sales organization necessitates empirical analysis. Nevertheless, the general relationship statement indicates that only salesman motivation and effort is under the control of the salesman in the short run. The other variables represent constraints under which he must operate.

**Territory performance.** Many criteria exist as possible indicators of territory performance. Sales volume (aggregate and by-product line) is frequently used in practice. Other criteria include product mix, number of new customers, number of orders, market share, profitability, and sales per customer. The territory goals considered important for a particular organization should be used as a basis for selecting one or more performance criteria.

**Market potential.** Market potential is "the capacity of a market to absorb a product or group of products of an industry in a specified period of time [Davis

& Webster, 1968, p. 259]." Actual industry sales are frequently used as an approximation of market potential. Alternatively, various indirect measures of potential can be used by identifying one or more factors which are correlated with market potential (e.g., number of employees or potential buyers in industrial markets).

**Workload.** The amount of effort necessary to generate the same volume of sales normally will be different in different territories. Workload is the input necessary to produce the same level of output (sales, profits, etc.) in all territories. Since the outputs of territories are typically not the same, it is logical to assume that workload or activity should account for a portion of the variation experienced in territory performance. Variations in workload are the result of (a) number, dispersion, tenure, and servicing requirements of customers and (b) physical characteristics of the territory (e.g., Rocky Mountain area compared to Manhattan).

**Salesman.** Both the salesman's experience ( $S_r$ ) and the motivation and effort ( $S_f$ ) he puts into his job are clearly relevant influences upon territory performance. Conceptually, these dimensions appear distinct. However, separating what the salesman has to work with (experience) from how he uses his capabilities and experience (effort) present difficult measurement tasks.

The salesman experience variable is a composite of the knowledge and skills possessed by the salesman. Experience is viewed as the capabilities of the salesman at a particular point in time and therefore, not controllable by him in the short run. There are many possible measures of experience including education, training, and job experience.

The salesman motivation and effort variable is an attempt to recognize the degree to which the salesman utilizes his capabilities. This factor is controlled by the salesman. The importance of motivation and effort is likely to vary depending upon the type of selling job involved (McMurray, 1961). Consider, for example, the differences in territory results that effort may represent in creative selling (such as life insurance) versus account servicing or order taking (such as delivery-sales jobs).

**Company.** The logic for separating company experience and effort is similar to that discussed relative to the salesman. Company standing (experience) is likely to differ by territory as is the amount of support (effort) provided by the company to a particular territory.

Company experience refers to the accumulated capabilities of the company in a given territory resulting from past efforts of salesmen, competitive strengths, length of time the territory has been open, management, etc. A logical composite or proxy measure of company experience is market share since it should be related to many aspects of company experience (both positive and negative).

Company effort is defined as the resources provided to a particular sales territory (promotion, information, home office support, etc.). These resources may assist the salesman and contribute to territory

TABLE 1  
SUMMARY OF VARIABLES AND MEASURES

Variable	Measure <sup>a</sup>
Criterion variable	Aggregate sales (in dollars) credited to territory salesman ( $TP$ )
Sales territory performance	Industry sales (in units) of products sold in territory as reported by the trade association serving the industry
Determinants of performance	Average workload per account using a weighted index based on annual purchases of accounts and concentration of accounts ( $W_1$ )
Market potential	Total number of accounts handled by salesman ( $W_2$ )
Territory workload	Length of time (in months) employed by company ( $S_c$ )
Salesman experience	Aggregate rating (1-7 scale) by applicable field sales manager on eight dimensions of performance ( $S_f$ )
Salesman effort	Weighted average of past market share magnitudes for 4 previous years ( $C_1$ )
Company experience	Market share change over 4 years previous to time period analyzed ( $C_2$ )
Company effort	Advertising expenditures (in dollars) in the territory ( $C_f$ )

<sup>a</sup> All measures are for the time period analyzed unless otherwise indicated.

performance. Examples of measures of company effort in the territory include amount of cooperative advertising, trade shows, and special pricing programs.

*Other influences.* These influences are likely to be specific to a particular firm. If considered sufficiently important, measures can be selected and included in the analysis. In most cases, the influences previously discussed should be sufficient as a basis for analyzing sales territory performance.

### Measurement of Variables

The identification of candidate variables to be included in the analysis was guided by the conceptual framework previously discussed. Actual selection of variables was based on extensive discussions with the firm's management. There was no indication that additional territory-specific factors were operating in the territories included in the study, so the "other influences" variable was excluded. The variables and corresponding operational measures selected for analysis are summarized in Table 1.

### Approach to Analysis

Multiple regression was used to analyze the relationships between the criterion and predictor variables. Other, more complex methods of analysis were considered. Yet, it was felt that, providing the method yielded acceptable levels of explanation of the variation in the empirical data, the availability of standard computer programs, coupled with its relative simplicity, made multiple regression an appropriate tool for use by practitioners.

### RESULTS

#### Preliminary Analysis

Analysis of data for the 25 territories using multiple linear regression yielded a coefficient

of multiple determination ( $R^2$ ) of .722. The strength of this empirical relationship was encouraging. Yet since market response relationships are frequently not linear, it seemed appropriate to further analyze the data using a curvilinear model.

The following curvilinear relationship was selected from a number of possible models since by transforming the data to logarithmic values, the analysis could be performed using the multiple linear regression analysis:

$$TP = A P^{b1} W_1^{b2} W_2^{b3} S_3^{b4} S_4^{b5} C_5^{b6} C_6^{b7} C_7^{b8}$$

The coefficient of multiple determination using the transformed data was .88 and was

TABLE 2  
SUMMARY OF STEPWISE MULTIPLE REGRESSION  
ANALYSIS USING DATA TRANSFORMED TO  
LOGARITHMIC VALUES

Step number	Variable entered <sup>a</sup>	Multiple R <sup>2</sup>
1	Length of employment	.730
2	Average market share	.809
3	Salesman performance rating	.871
4	Advertising expenditures	.876
5	Average workload per account	.878
6	Number of accounts	.879
7	Average market share change	.879
8	Industry sales	.879

<sup>a</sup> The predictor measures included in the regression relationship at a given step consist of the measure shown for that step plus measures shown for all previous steps.

significant at the .001 level. Thus, approximately 16% in additional explanation was obtained via the curvilinear model. Adjustment of the  $R^2$  for degrees of freedom resulted in a value of .83.

The results of a stepwise multiple regression analysis using the transformed data from the curvilinear model are shown in Table 2. Salesman experience, average market share, and salesman performance rating provided a major proportion of the explained variation in the criterion variable, aggregate sales in the territory.

#### *Predicted versus Actual Performance*

Given the strong relationship between the criterion and predictor variables, the final step in the methodology is to develop performance standards which have been adjusted for the relevant factors operating in each territory but which are not under the personal control of the salesman. A measure for salesman motivation and effort ( $S_7$ ) was included in the preliminary analysis in seeking to explain all variation in the criterion variable recognizing that it should not be included as an independent variable in determining salesman performance standards. (The eight dimensions of performance which made up  $S_7$  were: salesman's overall reputation in the industry; strength of his relationships with customers, and within the company; profitability of sales results; coverage of relevant markets; problem-solving effectiveness; quota performance; and sales development effort. These dimensions are the current company performance rating criteria.) This measure was removed and performance benchmarks (predicted values of the criterion variable using the regression equation) were calculated using only the two other major explanatory predictor variables, length of employment and average market share. Since the other predictor measures not controlled by the salesman (advertising expenditures, workload, number of accounts, market share change, and industry sales) did not contribute appreciably to the relationship, they were not included in the benchmark calculations. The coefficient of multiple determination for this relationship was .81 and was statistically significant at the

TABLE 3  
TERRITORY RANKINGS BASED ON SALES VOLUME,  
BENCHMARK ACHIEVEMENT, QUOTA ACHIEVEMENT,  
AND PERFORMANCE RATINGS

Territory number	Sales volume	Benchmark achievement	Management rating	Quota achievement
1	15	2	4	1
2	6	9	2	16
3	18	7	20	15
4	8	23	17	22
5	2	5	5	17
6	24	17	7	2
7	3	16	6	24
8	12	25	24	18
9	1	4	3	13
10	5	10	21	19
11	19	20	18	9
12	22	15	9	4
13	20	8	8	8
14	21	11	10	6
15	4	1	1	14
16	11	21	14	21
17	9	14	15	20
18	13	3	11	11
19	14	13	13	10
20	10	18	19	25
21	23	22	25	5
22	16	19	16	12
23	25	24	23	3
24	7	6	22	23
25	17	12	12	7

.001 level. Rankings of the 25 territories based on sales volume, benchmark achievement (sales volume divided by benchmark sales), management ratings of salesman motivation and effort, and quota achievement (actual sales divided by assigned sales quota) are shown in Table 3.

There is a definite mix of territories with both high and low absolute sales volumes included in those territories performing relatively well, based on the benchmark achievement rankings in Table 3. While this is not a validation of the benchmarks, the finding does meet with expectations. Under normal circumstances, relatively high salesman performance achievement would not necessarily be restricted to high absolute sales volume ter-

ritories. Since in this firm salesmen with longer tenure tend to have the territories with the highest sales, absence of high performance achievement in territories with lower sales would suggest either very poor new salesman selection procedures and/or an abnormally long period needed for learning the job. Neither of these phenomena appeared to be present in the sales organization.

An analysis was made of the relationship between the rankings of salesman performance ratings and rankings of (a) benchmark achievement and (b) quota achievement as shown in Table 3. The Spearman rank-correlation coefficient ( $r_s$ ) between benchmark achievement rankings and salesman performance rankings was .61 and was statistically significant at the .001 level (Siegel, 1956, pp. 202-213). Thus, salesman performance standards using the analytically determined benchmarks appear to be reasonably consistent with management's ratings of salesmen. Quota achievement, however, does not bear any strong relationship to management ratings ( $r_s = .17$ , not significant), and therefore, would not seem to be an appropriate measure of salesman performance. (Quotas were largely determined using market potential data and past results.)

## DISCUSSION

### *Role of the Salesman*

The analytically determined performance benchmark enables the salesman to perform against a yardstick that has been adjusted for the various differences that occur between sales territories and that are beyond his control. In the territories analyzed, the major contributing factor to explaining variations in sales was the length of time the salesman had been employed by the company. If the performance standards set by management are not adjusted for differences in tenure, then new men are likely to be confronted with inequitable gauges of their performance. An example of this is demonstrated by comparing benchmark achievement to quota achievement in Territory 3 (see Table 3). Only 2 of the 25 salesmen had less tenure with the company than the man in Territory 3. He ranked seventh in benchmark achievement compared

to fifteenth on quota achievement. Moreover, his performance ranking by management was near the bottom of the group. Based on benchmark achievement, a reevaluation of the man is indicated.

The results of the analysis suggest that in the organization analyzed the salesman is a necessary but not sufficient contributor to territory performance. Most important, the degree to which he can cause *major* increases in sales in the short run appears to be small in view of the uncontrollable factors related to sales results. If improvements in sales results are largely long term in nature, then it is extremely important that management determine sensitive gauges of performance in order to prevent young, competent men from becoming discouraged and leaving the company (the man in Territory 3 may well fall into this category). There are clear indications that the analytically determined benchmarks may be more helpful in this regard than the quotas previously used by the firm. Yet, management judgment and experience should play a central role in using the benchmarks for both analyzing past performance and predicting future performance.

### *Predicting and Improving Salesman Performance*

The approach provides a promising first step for studies attempting to predict salesman performance. Prior to selecting and testing predictor variables of interest, the methodology can be used to select appropriate criterion variables. For example, benchmark achievement (ratio of actual to predicted sales for each sales territory) appears to be one appropriate performance criterion for salesmen in the organization analyzed. The next step would be to analyze the relationship between benchmark achievement and variables related to the salesmen (personality, personal history items, attitudes, etc.).

The benchmark approach is quite flexible in filling the need for multiple performance criteria. Based on the goals established by management for the personal selling function, additional performance criteria can be analyzed in the same manner as was done for territory sales. After adjustment for factors beyond the

salesman's control, the resulting standards can be used in determining multiple performance benchmarks.

### Extensions of the Analysis

Since this research must be viewed as exploratory, several additional related avenues of study are indicated:

1. Determination of the stability of relationships over time through cross-sectional analysis of data for several past time periods.
2. Development of more sensitive measures for certain of the predictor variables (territory workload for example).
3. Application of the methodology using other performance factors (criterion variables) such as product mix, profitability, etc.
4. Replication of the methodology in other sales organizations particularly where the role and importance of the salesmen vary substantially.
5. Examination of the usefulness of the territory performance benchmarks in providing more sensitive criterion variables for use in combination with salesman characteristics to develop more effective guidelines for predicting salesman performance.

Many of the determinants of sales territory results are likely to be organization specific.

This suggests the need for a close relationship between the analyst and sales management. Identification of all possible predictor variables is a crucial aspect of the methodology. Salesman participation in the identification process may also yield valuable inputs to the analyst. This also could contribute to better acceptance of the resulting benchmarks as a basis for performance evaluation.

### REFERENCES

- BAEHR, M. E., & WILLIAMS, G. B. Prediction of sales success from factorially determined dimensions of personal background data. *Journal of Applied Psychology*, 1968, 52, 98-103.
- COTHAM, J. C., III. Job attitudes and sales performance of major appliance salesmen. *Journal of Marketing Research*, 1968, 5, 370-375.
- COTHAM, J. C., III. Selecting salesmen: Approaches and problems. *Michigan State University Business Topics*, 1970, Winter, 64-71.
- DAVIS, K. R., & WEBSTER, F. E., JR. *Sales force management*. New York: Ronald Press, 1968.
- MCMURRAY, R. N. The mystique of super-salesmanship. *Harvard Business Review*, 1961, 39, 113-122.
- MINER, J. B. Personality and ability factors in sales performance. *Journal of Applied Psychology*, 1962, 46, 6-13.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

(Received October 21, 1971)

## THE PERCEPTION OF ORGANIZATIONAL CLIMATE: THE CUSTOMER'S VIEW<sup>1</sup>

BENJAMIN SCHNEIDER<sup>2</sup>

*University of Maryland*

Climate was defined as the summary perception that bank customers have of their bank. Perceived climate was conceptualized as an intervening variable—a summary perception based on specific service-related events but preceding customer account switching. Questionnaire data obtained from 674 present and 87 former bank account holders indicated that (a) present customer intentions to switch accounts are more strongly related to summary perceptions than to specific service-related event perceptions of the bank and (b) former customers have significantly more negative perceptions of the bank and its employees than do present customers. Implications for future organizational climate research and for the relationship between employee and customer are discussed.

Most of behavioral marketing and consumer psychology research has been directed at attracting consumers to products. There has not been much research on the retention of product consumers and little if any research on the attraction and retention of the consumer of services. In a review of 178 consumer psychology articles (Twedt, 1965), there were no studies of service organizations. Popular texts in industrial psychology (Blum & Naylor, 1968; Tiffin & McCormick, 1965), readings and books in the same general area (e.g., Fleishman, 1967), and readings in consumer behavior (Kassarjian & Roberts, 1968) have failed to consider the attraction and retention of the consumer of services.

The present research develops a framework for beginning to understand some of the bases of the global perceptions people have of organizations. The role of these perceptions as correlates of customer account switching is explored in a bank setting. The following hypothesis is investigated: In service organizations characterized by employee-customer face-to-face contact, customers have summary perceptions about organizations that may be based on

their perceptions of specific service-related events and behaviors. Where external forces (e.g., a permanent move demanding a switch) are not a factor, summary perceptions (i.e., perceived organizational climate) may be a basis for customer decisions to remain with or leave the bank.

There are a number of assumptions underlying this hypothesis. First, it is assumed that service organizations are open systems that interact with, influence, and are influenced by segments of the society in which they exist. Thus, the way employees behave toward customers is thought to be the result of the work climate that the bank creates for them; employees, in turn, create the climate that the customers perceive. Some support for this assumption has been presented by Pickle and Friedlander (1967) who showed that across 97 small organizations, employee and customer satisfaction were significantly intercorrelated ( $p < .05$ ). They further demonstrated that the ability characteristics of the organization manager, especially critical thinking skills, were related to both employee and customer satisfaction. Second, the proposed framework assumes certain characteristics of people and the perceptual process; that is, customers have perceptions of specific events and behaviors in organizations that they may use as a basis for formulating their summary perception. The summary perception concerning the larger organization is defined as perceived organizational climate.

An individual difference component in the

<sup>1</sup>The author would like to thank Matthew Sobel for his help in carrying out this research project. H. Peter Dachler, Gerritt Wolf, Joseph Schneider, Douglas T. Hall, Chris Argyris, Edward E. Lawler, Irwin L. Goldstein, and Sharon B. Dorfman all provided useful comments on earlier versions of this article. Special thanks are due the management of the bank for providing the opportunity to conduct this research.

<sup>2</sup>Requests for reprints should be sent to Benjamin Schneider, Department of Psychology, University of Maryland, College Park, Maryland 20742.

present research takes the form of situation-specific values, that is, those aspects of the relationship between the individual and organization to which individuals attach importance. The relative value of organizational characteristics to individuals may be related to the events individuals perceive in the organization as well as the climate perceptions formed from these perceived events. If the occurrence of particular events is important to an individual, he is more likely to note that the event has occurred. Furthermore, the resultant climate perceptions should reflect the events and the event perceiver. Thus, it is further hypothesized that customer situation-specific values may be related to the reported occurrence of events and the organizational climate perceptions.

## METHOD

### *Research Sites*

Four representative branches of a prominent Northeastern commercial bank were selected for study; two branches were primarily retail (nonbusiness) and two predominantly commercial. The two commercial branches had more account holders, larger bank balances, more visitors per day, and shorter average customer waiting time than the retail branches.

### *Pilot Questionnaires*

As the result of 15 to 20 personal interviews in each branch bank, six bank features were found to be important to customers: (a) convenience, (b) short waiting time, (c) personal friendly service, (d) full-service banking, (e) safety, and (f) decoration. Categories *a* through *d* were mentioned by the customers themselves, while *e* and *f* were mentioned explicitly by bank personnel as characteristics they thought were important to customers.

Items descriptive of each of the above six categories and items designed to tap customer intentions to switch accounts to another bank were written. After the first set of questionnaires was administered to 275 customers, necessary wording changes were made. The revised set of pilot questionnaires was administered to a new group of 284 customers. All pilot questionnaires were collected in the branches. Items from both administrations that were significantly correlated with the behavioral intention to switch accounts were retained for the final questionnaire.

The directions for responding to the items were adapted from the Job Description Index (Smith, Kendall, & Hulin, 1969). Customers responded with a "y" (for Yes if it describes your experiences or feelings), "n" (for No if it does not describe them), or "?" (if you cannot decide) to each item. The y, ?, and n were scored 3, 2, and 1, respectively.

Data on actual and perceived waiting time from the moment a customer joined a queue to completion of transaction were collected for a portion of both the pilot questionnaire samples (total  $N = 305$ ) to assess the impact of immediate situational contaminants on item responses. Actual and perceived waiting time, in minutes, correlated highly ( $r = .81$ ,  $p < .01$ ), but neither was related to questionnaire item responses.

### *Final Questionnaire*

The final questionnaire contained 13 items selected from the two pilot questionnaire administrations. In addition to these items, there were questions related to (a) the type(s) of account(s) the customer had, (b) how often the customer visited the bank, (c) sex, (d) marital status, (e) the distance the customer lived or worked from his branch, "whichever is closer," and (f) the importance of various bank features or services. The latter was assessed in the following manner:

We need to know how important different bank features are to you, regardless of the way they are now. Check the list below and if a feature that is important to you does not appear, add it to the list. Now pretend someone has given you exactly \$100 to spend on one or more of the features. The hitch is you must spend it all, and only on the features. Simply indicate on the line beside the feature how much of the \$100 you are willing to spend on it.

The six features were then listed and customers were asked to check that the total spent was \$100.

### *Final Samples*

From each branch, a sample of 165 names was randomly drawn from current listings of the most used bank services: savings accounts, regular checking (usually a commercial account), and special checking (usually a personal account). The sample was obtained by taking the total number of account holders of each type and dividing this number by 165. The quotient indicated the interval to be used between names in selecting the sample. In addition to names and addresses, account type and account balance were also noted on the address label. Account balance was coded by a 1-9 scale appropriate to the type of account being coded. Questionnaires were similarly coded.

Of the 1,980 questionnaires mailed to current customers, 674 (34%) were returned. Estimated account balances for the people who returned questionnaires suggest that they are a representative sample of the entire mailed sample (see Table 1). Across the four branches there are no significant differences between estimated account balances for the total mailed sample and the returned sample. However, as noted earlier, the commercial branch depositors hold larger bank balances, especially under regular checking.

A sample of former account holders was also sent questionnaires with all questions worded in the past tense: "Think of the \_\_\_\_ branch where you banked. How well does each of the following statements describe

TABLE 1

COMPARISON OF AVERAGE ACCOUNT BALANCES FOR THOSE RECEIVING QUESTIONNAIRE  
AND THOSE RETURNING QUESTIONNAIRE

Type of account	Branch			
	Retail 1	Retail 2	Commercial 1	Commercial 2
Special checking				
Questionnaire received	473.21	356.06	475.74	465.43
Questionnaire returned	452.00	358.45	676.85	519.79
Regular checking				
Questionnaire received	1050.30	1838.34	3113.64	4645.45
Questionnaire returned	972.97	1755.21	3814.66	5629.31
Savings				
Questionnaire received	384.45	421.82	424.55	495.56
Questionnaire returned	337.88	453.49	690.91	392.86

your experiences or how you felt about banking at your branch?" On these questionnaires, an additional item, "I closed my account because . . ." was included so that customers who switched their account because of physical relocation, business failure, etc. could be separated from those who switched for reasons over which the bank may have control.

One hundred and twenty-one of 600 questionnaires (20%) sent to former account holders were returned. Sixty-three returns had closed their accounts because of a move, going out of business, death, or retirement; this group was used as a control group against which to compare the 24 returns who switched because of service issues. Thirty-four people were dropped from

TABLE 2  
CORRELATION OF ITEMS WITH SWITCH TENDENCY BY BRANCH

Item	Retail 1 (n = 125)	Retail 2 (n = 172)	Commercial 1 (n = 202)	Commercial 2 (n = 175)
1. I depend on the bank for all kinds of banking services.	.07	-15	-05	-15
2. I try to use the same teller each time I bank.	.20	.02	-15	-02
3. It doesn't seem like the tellers help each other out when the bank gets busy.	.12	.25	.33	.18
4. High caliber people work in the bank.	-.24	-.20	-.30	-.20
5. The bank employees bend over backwards to provide good service	-.41	-.40	-.35	-.28
6. My branch is the most convenient place for me to bank.	-.30	-.15	-.04	.02
7. I have to wait in line too long at the bank.	.41	.33	.20	.30
8. Things happen in the bank that make me want to switch my account elsewhere.				
9. The branch employees seem happy about the fact that they work here.				
10. The bank employees treat all customers as equals.	-.27	-.28	-.22	-.23
11. I get so irritated at my bank that I think of switching my account.	-.41	-.25	-.17	-.11
12. The atmosphere in my bank is warm and friendly.				
13. When I get to the bank I first look to see which of the lines is longest.	-.48	-.51	-.35	-.33
Correlations of switch items with each other	.12 83	.02 75	-.02 72	.04 65

Note. Italicized words appear in subsequent tables. Correlations indicate the average correlation with Items 8 and 11. Decimal points have been omitted.

further analyses because they had not closed their account, had changed their account name through marriage or merger, or had moved and switched their account to another branch. Several other respondents were not included because they failed to indicate their reason for closing or did so in an ambiguous fashion (e.g., personal reasons).

## RESULTS

Analyses supporting the pooling of data from account maintainers from the four different branches are presented in Table 2. These data are viewed as four replications of the correlation of the items and switching intentions for current account holders. Items descriptive of employee behavior tended to have the highest correlations with customer intentions to switch. The strongest correlate of switch tendency in all branches is Item 12—the summary perception, "the atmosphere in my bank is warm and friendly." In Table 2 and all subsequent tables, correlations involving switch intentions are the average of the correlations with Items 8 and 11.

The average level of item responses as well as the item-switch correlations are highly similar across branches, so subsequent analyses are accomplished on the pooled sample ( $N = 674$  maximum). The three exceptions to this generalization are all in Retail 1 and concern Items 2 (using the same teller), 6 (most convenient branch), and 13 (I look to see which of the lines is longest). Retail 1 customers, relative to customers in the other branches, do not attempt to use the same teller, do not look to see which line is longest, and do feel their bank is the most convenient (all  $p < .01$ , two-tailed tests). In this branch, apparently because the system for queuing customers results in only one line, customers are not able to choose their teller. In addition, this branch is the only bank in a six-block area that offers checking accounts; it is literally more convenient than any other.

### *Perception Correlates of Account-Switching Intention*

Table 3 presents the intercorrelations of bank perceptions, service values, and switch intentions. The data in italics indicate correlations with the switch tendency items.

It seems clear in Table 3 that perceptions

by customers of the bank, as in interpersonal employee-customer relationships, are important as correlates of switching intentions. Except for Item 7 (I have to wait in line too long at the bank), only perceptions concerning the interpersonal nature of the employee-customer relationship and/or the employees themselves are appreciably related to switching intentions. Thus, Items 1 (depend on bank), 6 (most convenient place), and 13 (see which line is longest) are not related to switching intention. In addition, none of the more objective indexes of customer participation such as bank balance, distance from the bank, length of time as bank customer, etc., were appreciably related to account-switching intentions.

Further analysis of the relationships between customer perceptions and customer intentions to switch their accounts suggest that the correlates of switch intentions can be grouped into two sets. One set contains Item 12 (atmosphere is warm and friendly;  $r = -.41$ ,  $p < .01$ ) and Item 5 (employees bend over backward;  $r = -.38$ ,  $p < .01$ ). These two items seem to define a summary or climate perception.

The second set of items had consistently and significantly lower correlations with switch intentions ( $r = [.23]$ ): Item 3 (tellers do not help each other), Item 4 (high caliber people), Item 9 (employees seem happy), and Item 10 (employees treat all as equals). These items seem to be more specific perceptions of service-related events than are Items 5 and 12.

The cluster of specific perceptions is (a) not as strongly related to switch intentions as the summary perceptions are ( $r = [.23]$ ,  $r = -.40$ , respectively;  $t = 2.63$ ,  $p < .01$  [McNemar, 1962]) and (b) more strongly related to the cluster of climate perceptions than to the cluster of switch intentions ( $r = [.40]$ ,  $r = [.23]$ , respectively;  $t = 2.63$ ,  $p < .01$  [McNemar, 1962]).

These data suggest tentative support for the hypothesis that summary perceptions may be based on more specific perceptions and that customer behavior may be based more on summary perceptions than on less abstract perceptions. In any case, it seems clear that perceptions of the bank as an interpersonal employee-customer relationship is important as a correlate of switching intentions. Item 7

TABLE 4  
MEANS, STANDARD DEVIATIONS, AND *t* TESTS FOR TOTAL SAMPLE, SWITCHED SAMPLE,  
AND CONTROL SAMPLE

Item	Total sample		Switched sample		Control sample		<i>t</i>	
	(n = min 582)		(n = min 20)		(n = min 52)			
	<i>M</i>	$\sigma$	<i>M</i>	$\sigma$	<i>M</i>	$\sigma$		
1. Depend on the bank.	2.08	.98	2.29	.99	2.18	.98	1.03	
2. Use the same teller.	1.48	.85	1.33	.78	1.53	.90	.85	
3. Tellers do not help each other.	1.91	.85	2.45	.69	1.86	.83	-3.07*	
4. High caliber people.	2.27	.74	1.50	.52	2.38	.75	5.04*	
5. Employees bend over backwards.	2.10	.85	1.54	.88	2.28	.84	3.15*	
6. Most convenient place.	2.43	.88	2.38	.96	2.60	.81	.27	
7. Wait in line too long.	2.30	.91	2.54	.88	1.89	.97	-1.27	
8. Things happen that make me want to switch.	1.52	.85	2.31	.95	1.29	.71	-4.44*	
9. Employees seem happy.	2.41	.63	1.92	.79	2.47	.63	3.70*	
10. Employees treat all as equals.	2.59	.67	2.25	.96	2.80	.41	2.40*	
11. I get irritated at my bank.	1.41	.77	2.38	.96	1.04	.20	-5.98*	
12. Atmosphere is warm and friendly.	2.48	.74	1.54	.88	2.53	.73	6.05*	
13. See which line is longest.	2.66	.74	2.50	.90	2.52	.87	1.02	

Note. *t* test is for total sample versus switched sample; total sample includes those who intend to switch.  
\*  $p < .01$ , one-tailed test.

(wait in line too long) does not fit the interpersonal mold, yet it is correlated significantly with switch intentions ( $r = .30, p < .01$ ).

#### Service Value Correlates of Switching Intention

The data in Table 3 indicate that customers who more highly value personal, friendly service perceive the bank and its employees in positive terms (warm and friendly, high-caliber employees), while those valuing very short waiting time perceive the bank and its employees more negatively. The strongest correlate of both situation-specific service values is Item 7, wait in line too long ( $r = .31$ ) for the importance of personal, friendly service. The correlations of the other items with the service values are all .20 or lower and concern the specific perceptions and climate perceptions referred to earlier.

Situation-specific service values were marginally related to switch intention, with the highest correlation being .13. Thus, while the situation-specific values are related to specific and summary perceptions, they are not meaningfully related to switch intention.

#### Former Account Holders

Table 4 presents data for the samples of present account holders, those who switched their accounts for reasons over which the bank may have control, and those who switched their account for reasons such as a physical move. Except for Item 7, waiting in line too long, the data indicate that all of the items that correlated with switch tendency also discriminated between those who maintain and those who switch their accounts. The *t* values for the two climate perceptions (Items 5 and 12) are 6.05 and 3.15 for an average of 4.60. For those *t* values related to specific perceptions (Items 3, 4, 7, 10) only Item 4, high-caliber people work in the bank, is higher than 4.60 (*t* = 5.04). The average *t* for the specific perceptions was 3.54.

These data suggest further support for the hypothesis that customer perceptions of organizational climate are related to customer account-switching behavior. The magnitude of the *t* values (given equivalent sample sizes) indicates that specific perceptions are not as strongly related to account switching as are the summary climate perceptions.

Tests of significance calculated on the situation-specific service values yield one significant  $t$ . Personal, friendly service is less important ( $\bar{X} = 11.18$ ,  $\sigma = 11.11$ ) for customers who switched their accounts than for customers maintaining their accounts ( $\bar{X} = 17.68$ ,  $\sigma = 19.77$ ,  $t = 2.26$ ,  $p < .05$ ). This finding would not have been predicted on the basis of the correlational data presented for account maintainers.

### DISCUSSION

The present research has shown that bank customer intentions to switch their accounts are significantly related to their perceptions of bank employees and the climate of the bank. In addition, the data indicated that account holders who had switched their accounts for service-related reasons had perceptions of the bank and its employees that were significantly more negative than the perceptions of customers still maintaining their accounts.

The results give some tentative support to the following generalizations: first, climate perceptions of an organization (e.g., how warm and friendly it is) may be summary perceptions of events or experiences perceived by people who interact with it. Customer perceptions of the bank's climate may be based on customer perceptions of bank employees—the perceived caliber of employees, whether employees help each other in serving customers, whether employees treat all customers as equals, and whether employees seem happy in their work.

Second, people may leave an organization as a result of their summary perceptions of the organization. Climate perceptions are more strongly related to switching behavior than are the perceptions of specific events and experiences.

Third, perceiver situation-specific service values are consistently, but not strongly, related to both specific event and summary climate perceptions. Although these values are not strongly related to behavior intentions, they can be related to actual behavior. Specifically, customers who actually did switch, rated the importance of personal, friendly service significantly lower than customers who remained.

Fourth, more objective characteristics of customers (size of account, type of account,

distance from bank, length of time with bank, sex, number of bank services used) and of the bank (actual waiting time, size of accounts, procedure for queuing customers) were unrelated to specific event and summary climate perceptions.

Given the above generalizations, there are two major issues to be discussed: the methodology and conceptualization of climate and the relationship between employee and customer.

#### *Climate: Concept and Method*

When a psychologist assumes that individuals behave in or toward organizations on the basis of their global perceptions, he also assumes that a correlate of behavior in organizations is individual attributes. This assumption dictates a micro approach to predicting behavior, involving such data as individual ability, needs, and values. On the other hand, work group performance or organizational turnover may be the focus of interest. In that case, the nature of the group or organizational task (Dubin, 1968), the group or organizational reward system (Campbell, Dunnette, Lawler, & Weick, 1970), etc., dictate a macro approach to understanding climate perceptions.

The concept of climate in the present research may best be described as personalistic; climate is an individual perception. There was no attempt to restrict the climate definition to perceptions shared by members of a work group or organization. As stated elsewhere (Schneider & Bartlett, 1970), ". . . what is psychologically important to the individual must be how he perceives his work environment, not how others might choose to describe it [p. 510]." Perhaps, however, shared perceptions are important when predicting the behavior of many individuals. Individual perceptions may only be important for predicting individual behaviors. The researcher must be clear about the level of his research question (Weick, 1968), so that the data collected corresponds to the level of the phenomenon being predicted.

The major methodological contribution of the present research to the study of climate is a definition of climate tied to behavior in both the independent and dependent variables: the behavior of bank employees and the behavior of bank customers. Conceived of in

this way, climate perceptions are intervening variables caused by discrete experiences and causing later behaviors (Likert, 1961). The important concept is that people may perceive specific elements in a situation that may in turn be related to a summary perception of the organization. This summary perception may serve as a basis for behavior toward the organization. Because relationships are specified in dynamic terminology, this is a processual view of climate, one that dictates the collection of data over time. Thus, on the basis of the present study, one may only speculate about causal sequences. Further research on the development or emergence of climate perceptions is clearly required (Beer, 1971).

The framework proposed in the present research suggests, for example, that the longer individuals have been in contract with an organization, the more difficult it will be to affect their climate perceptions. Over time, as the result of many specific perceptions, the summary perceptions that constitute the individual's conception of an organization climate should become less subject to change. It follows, then, that early in an individual's association with an organization a perception of specific events may have more of an effect on the summary perceptions than the same perception would have at a later time period. This might account for the reported tendency of climate perceptions to remain consistent over time (Greiner, Leitch, & Barnes, 1968), for the difficulties encountered in bringing about new climate perceptions (Beer, 1971), and for the unusually high impact early experiences in organizations have on later performance (Berlew & Hall, 1966; Hall & Schneider, 1973).

#### *Employees and Customers: The Insider-Outsider Relationship*

Behavioral scientists do not know whether the impact of formal organizations on organizational members extends beyond the organization. Katz and Kahn (1966) assume that people outside formal organization boundaries cannot understand what happens inside the organization. Perhaps service organizations are a special case of open systems, that is, organizations in which the reason for existence is to serve outsiders.

In the present research it was assumed that the climate bank employees create for customers is an extension and result of the climate bank management creates for employees. It follows, then, that if data were collected on variables such as the satisfaction of bank employees, customer perceptions of how happy employees are should be correlated with employee reports. As noted earlier, Pickle and Friedlander (1967) have provided support for this hypothesis.

Perhaps information about organizational characteristics does permeate the boundaries of organizations. Pruden (1969) notes that Chester Barnard (1948) argued for the inclusion of the customer in the social system boundaries of business organizations. If the assumption of boundary permeability between the server and the served is a viable concept, then the application of the climate concept proposed earlier in this article has an important utility component. The framework enables the researcher to predict perceiver behavior and may also permit identification of the specific elements of the situation on which the climate perceptions are based. By examining these elements, changes that may result in altered climate perceptions can be specified.

If employee behavior toward customers is the result of climate created for employees and if customer behavior is the result of climate created by employees, the chain of events resulting in some customer account switching is identifiable. Applying the specific event perceptions, summary climate perceptions framework, to both employees and customers might provide an understanding of one of the underlying factors in customer account-switching behavior.

#### REFERENCES

- BARNARD, C. I. *Organization and management*. Cambridge, Mass.: Harvard University Press, 1948.  
BEER, M. Organizational climate: A viewpoint from the change agent. Paper presented at the meeting of the American Psychological Association, Washington, D.C., September 1971.  
BERLEW, D. E., & HALL, D. T. The socialization of managers: Effects of expectations on performance. *Administrative Science Quarterly*, 1966, 11, 207-223.  
BLUM, M. L., & NAYLOR, J. C. *Industrial psychology: Its theoretical and social foundations*. New York: Harper & Row, 1968.  
CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E.,

- & WEICK, K. E. *Managerial behavior, performance and effectiveness*. New York: McGraw-Hill, 1970.
- DUBIN, R. (Ed.) *Human relations in administration*. (3rd ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1968.
- FLEISHMAN, E. A. (Ed.) *Studies in personnel and industrial psychology*. (Rev. ed.) Homewood, Ill.: Dorsey, 1967.
- GREINER, L. E., LEITCH, D. P., & BARNES, L. B. The simple complexity of organizational climate in a governmental agency. In R. Taquini & G. Litwin (Eds.), *Organizational climate: Exploration of a concept*. Boston: Harvard Business School, Division of Research, 1968.
- HALL, D. T., & SCHNEIDER, B. *Organizational climates and careers: The work lives of priests*. New York: Seminar Press, 1973.
- KASSARJIAN, H. H., & ROBERTS, T. S. (Eds.) *Perspectives on consumer behavior*. Glenview, Ill.: Scott Foresman, 1968.
- KATZ, D., & KAHN, R. L. *The social psychology of organizations*. New York: Wiley, 1966.
- LIKERT, R. *New patterns in management*. New York: McGraw-Hill, 1961.
- MCNEMAR, Q. *Psychological statistics*. (3rd ed.) New York: Wiley, 1962.
- PICKLE, H., & FRIEDLANDER, F. Seven societal criteria of organizational success. *Personnel Psychology*, 1967, 20, 165-178.
- PRUDEN, H. O. The outside salesman: Interorganizational link. *California Management Review*, 1969, 12, 57-66.
- SCHNEIDER, B., & BARTLETT, C. J. Individual differences and organizational climate, II: Measurement of organizational climate by the multitrait-multirater matrix. *Personnel Psychology*, 1970, 23, 493-512.
- SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes*. Chicago: Rand McNally, 1969.
- TIFFIN, J., & MCCORMICK, E. J. *Industrial psychology*. (5th ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1965.
- TWEDT, D. W. Consumer psychology. *Annual Review of Psychology*, 1965, 16, 265-294.
- WEICK, K. Experimentation in translevel interaction. In B. P. Indik (Ed.), *People, groups, and organizations*. New York: Teachers College Press, 1968.

(Received December 22, 1971)

## PERFORMANCE EFFECTIVENESS AND EFFICIENCY UNDER DIFFERENT DYADIC WORK STRATEGIES<sup>1</sup>

SAMUEL C. SHIFLETT<sup>2</sup>

*United States Army Medical Research Laboratory, Fort Knox, Kentucky*

The effects of three task-solving strategies on group efficiency and effectiveness were studied. Sixty soldiers worked in dyads under a shared labor strategy or one of two divided labor strategies. Tasks included a difficult and an easy crossword puzzle. On the average, dividing labor resulted in greater efficiency (amount of work per man hour) while requiring subjects to work together resulted in substantially greater group effectiveness (total performance), but this effect occurred primarily on the easy task. It was suggested that a high degree of member interdependence maximizes redundancy of task-relevant abilities, resulting in generally superior performance effectiveness but frequently at the cost of efficiency.

It has been argued that groups can potentially increase performance through redundancy of ability (Davis, 1969; Zajonc & Smoke, 1959). That is, if a task requires all group members to work together and if individual performance is such that some probability of failure to perform adequately exists, then redundancy of ability or task relevant knowledge increases the probability that the task will be performed adequately. Furthermore, as a task becomes more difficult the probability of performance failure presumably increases, and so in order to maintain adequate group performance, the necessity for redundancy also increases. On a very easy task the necessity for redundancy disappears since the probability of an individual failing, or making an error, approaches zero.

The amount of redundancy in a group can be manipulated in several ways, as suggested by Goldman (1965), Laughlin, Branch, & Johnson (1969), and Steiner (1966). Shiflett (1972) attempted to manipulate redundancy by varying member interdependence. He found that variations in dyadic organizational structures resulted in different levels of performance efficiency and effectiveness. The term efficiency refers to group productivity in terms of man hours (Taylor & Faust, 1952) while effectiveness refers to maximum group performance, without regard to time. This distinction is

similar to the distinction made between speed and power in ability testing. Shiflett (1972) found that a shared labor organization, where both members were required to work together, and therefore a high redundancy situation, resulted in greater effectiveness but somewhat less efficiency than a divided labor strategy, where redundancy was effectively nil, on both an easy and a difficult task. These findings were in contrast to the hypotheses that divided labor would be equally effective on an easy task, as well as more efficient, and that shared labor would be more efficient and more effective when the task was difficult. Failure to fully support these hypotheses was attributed to the particular manner in which the labor was divided. Divided labor groups solved crossword puzzles in which one member had only vertical definitions, the other had only horizontal definitions, and the two members were not permitted to discuss their definitions with each other. This particular division of labor introduced communications and feedback difficulties by introducing relatively high task interdependence with low content-related communicability. If one member made an error it became more difficult for the other member to fill in his adjoining words and the restriction on communication made it difficult for members to locate the error. This was particularly true since each member had no way of determining whether he had made an error on the basis of his own performance; he could do this only through vague communication with his partner. A more appropriate labor division that would eliminate these problems would be to

<sup>1</sup> The author thanks James H. Davis for commenting on an earlier version of this article.

<sup>2</sup> Requests for reprints should be sent to Samuel C. Shiflett, Army Research Institute, Room 239, Commonwealth Building, 1300 Wilson Boulevard, Arlington, Virginia 22209.

allow each member to work on one intact half of each puzzle.

The purpose of this study was to replicate portions of the Shiflett (1972) study incorporating the appropriate modifications mentioned above. It was expected that on an easy task, the modified divided labor strategy would be more efficient than the shared labor strategy because of the reduced redundancy, and that it would be equally effective because redundancy was not necessary. On a more difficult task the shared labor strategy was expected to be more effective and more efficient than the modified divided labor strategy because of the necessity for increased redundancy. The modified division of labor was expected to be superior to the original vertical-horizontal division of labor in both efficiency and effectiveness.

## METHOD

### Subjects

Subjects were 60 soldiers who had recently completed basic training. The men were assigned to the research laboratory for 6-week periods in groups ranging from 16 to 20 men. The experiment was conducted during the second or third week of their duty at the laboratory, and the men within each group were acquainted with one another prior to participation in this experiment. The men ranged in age from 18 to 24, and in education from less than a high school diploma to college graduate. Although men scoring below 100 on the Army GT test were never assigned to the laboratory, the mean puzzle-solving ability of the soldiers, as assessed by the pretest described by Shiflett (1972), was more than one standard deviation below the ability of the college population used in the 1972 study.

### Task

Two crossword puzzles, one relatively difficult and one relatively easy, were cast in a symmetrical "skeleton" design in which each word had either one or two letters which were not shared with any other word. Each puzzle contained 48 four-letter words. No words were repeated within or across puzzles. While subjects worked on the puzzles, the experimenter observed them with a scoring sheet containing a copy of the puzzle outline. Whenever a word was written into the puzzle, the experimenter entered the time into the corresponding location of his own puzzle outline. Groups worked on each puzzle for 20 minutes. The number of correct words filled in during each 2-minute block was then tabulated, yielding word frequencies for each of the 10 blocks during the 20 minutes. Half of the dyads worked the easy puzzle first, and half worked the difficult puzzle first. At the end of each session subjects filled out a short questionnaire consisting of a series of bipolar scales assessing activity and satisfaction.

TABLE 1

SUMMARY OF ANALYSIS OF VARIANCE  
OF PERFORMANCE

Source	df	MS	F
Between subjects			
Order (O)	1	5.23	<1
Strategy (S)	2	52.08	.519*
O $\times$ S	2	1.31	<1
Error (Between)	24	10.04	1-
Within subjects			
Time (T)	9	204.01	47.29***
T $\times$ O	9	2.14	<1
T $\times$ S	18	20.53	4.76***
T $\times$ O $\times$ S	18	4.43	1.03
Error (T $\times$ subjects within)	216	4.31	
Difficulty (D)	1	154.03	177.59***
D $\times$ O	1	1.13	1.30
D $\times$ S	2	5.89	6.79**
D $\times$ O $\times$ S	2	1.73	2.00
Error (D $\times$ subjects within)	24	.87	
T $\times$ D	9	8.29	2.55**
T $\times$ D $\times$ O	9	2.19	<1
T $\times$ D $\times$ S	18	3.06	<1
T $\times$ D $\times$ O $\times$ S	18	1.75	<1
Error (TD $\times$ subjects within)	216	3.25	

\* p &lt; .05.

\*\* p &lt; .01.

\*\*\* p &lt; .001.

### Procedure

Subjects were randomly paired and assigned to one of three organizational strategy conditions. Subjects always worked on both puzzles using the same labor strategy. The first two conditions described below were identical to their counterparts described by Shiflett (1972). The third condition was the modified divided labor strategy.

#### Shared Labor Strategy

Subjects were given a single puzzle outline and a single set of definitions. They were told that they must work together on each word in the puzzle and must both agree on a word before writing it down.

#### Vertical-Horizontal Division of Labor Strategy

The experimenter placed a single puzzle outline between the subjects and explained that one of them would work only the horizontal words and the other only the vertical words. Each subject then received his set of definitions. Subjects were allowed to converse as much as they wished, but they could not indicate to each other what was printed on their own definition sheet.

#### Diagonal Division of Labor Strategy

This condition was identical to the vertical-horizontal division with the following exception. The puzzle outline had a line drawn diagonally through the puzzle, dividing the outline into two equal parts. The experimenter placed the puzzle outline between the subjects and explained that one of them would work only the words in the area above the diagonal and the other only the words below the diagonal. Each subject then received the appropriate set of definitions. They were

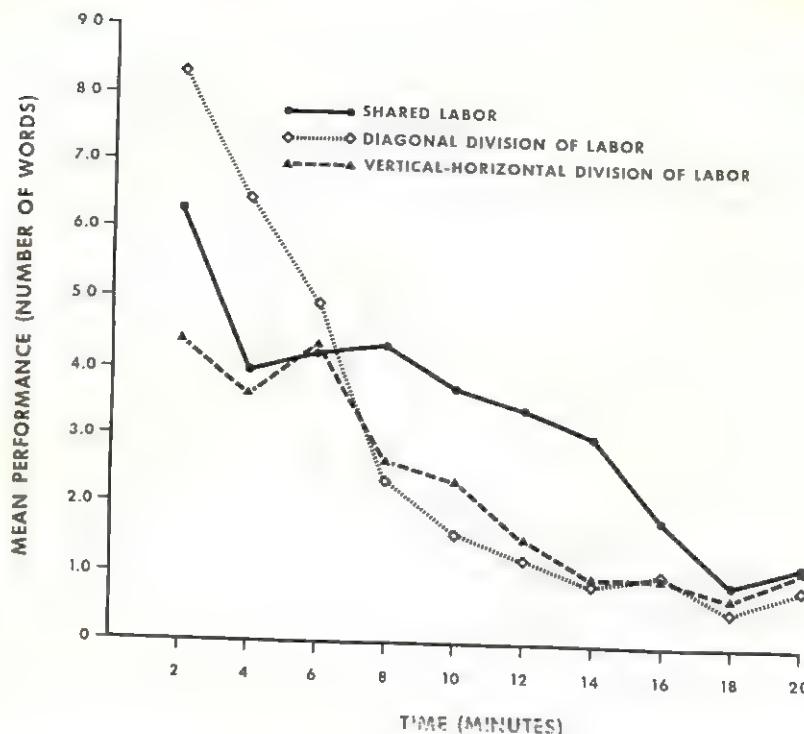


FIG. 1. Mean group performance across time for three different labor strategies.

allowed to talk to each other but could not discuss the definitions.

### RESULTS

The number of words completed per 2-minute period was calculated for each group, and constituted the measure of group performance. These data were submitted to a 4-way analysis of variance with repeated measures on two factors. The summary of this analysis is presented in Table 1. Effects of Time, Difficulty, and Strategy were significant, as were all three 2-way interactions involving these factors. On the average, more words were correctly completed on the easy puzzle per 2-minute block than on the difficult puzzle (3.27 vs. 2.26). The mean number of words per 2-minute block declined significantly from a high of 6.31 during the first 2 minutes to 1.06 words during the last 2 minutes suggesting that the tasks became more difficult as work progressed. Shared labor produced the highest level of performance with an average of 3.27 words per block; vertical-horizontal division of labor produced the lowest level of performance with an average of 2.25 words per block;

the diagonal division of labor was intermediate in performance with 2.79 words per block. The studentized range statistic indicated that each of these three means was significantly different from the others at the .01 level. This result thus substantiated the hypothesis that dividing labor vertically and horizontally produced poorer group effectiveness than a diagonal division. However, contrary to the prediction that the shared labor and diagonal division of labor would be equally effective was the finding that shared labor was significantly more effective than either of the divided labor strategies.

The Time  $\times$  Difficulty interaction indicated that performance on the easy puzzle was significantly greater than on the difficult puzzle during the first 6 minutes but not during the subsequent 14 minutes. The Strategy  $\times$  Difficulty interaction indicated that on the easy task, shared labor performance was significantly greater than performance under either of the divided labor strategies; whereas on the difficult task, shared labor performance and diagonal division of labor both exceeded verti-

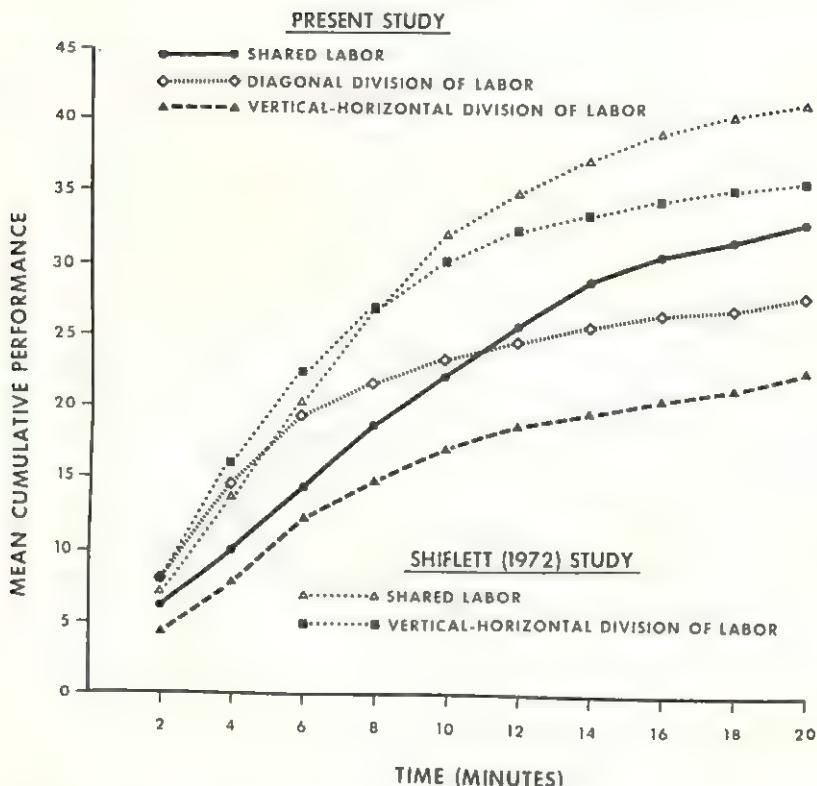


FIG. 2. Cumulative group performance for each of the labor strategies in the present study and for the corresponding strategies from the Shiflett (1972) study.

cal-horizontal division of labor, but did not differ from each other. In other words, the shared labor strategy resulted in greater effectiveness than divided labor on the easy task but not on the difficult task, thereby contradicting the basic hypothesis regarding the interaction between strategy and task difficulty.

The Strategy  $\times$  Time interaction, shown in Figure 1, indicated that during the first 6 minutes, diagonal division of labor yielded better performance than shared or vertical-horizontal division of labor, while after 8 minutes, shared labor performance exceeded that of both divided labor conditions. The vertical-horizontal divided labor performance generally paralleled shared labor performance during the first 6 minutes but closely paralleled diagonal division performance from minute 8 to 20. The significant differences between shared labor and diagonal division of labor and the change in the sign of the mean differences constitute support for the hypothe-



Time-to-criterion scores were obtained to test the hypothesis that when performance effectiveness was equated, divided labor would be more efficient than shared labor. Performance on the difficult task was at such a low

level that an analysis of time data for this task was not attempted. On the easy task, a criterion of 25 words<sup>3</sup> was used, requiring that two groups from each of the three strategy conditions be dropped from the analysis. In addition, times to subcriteria (5, 10, 15, and 20 words) were obtained and an analysis of variance containing two factors—Strategies and Criteria—was performed on the time scores. The summary of this analysis is presented in Table 2. Diagonal division of labor was the most efficient organization requiring 6.75 minutes to reach criterion while vertical-horizontal division of labor was least efficient using 14.35 minutes to reach criterion. Shared labor was intermediate in efficiency, requiring 9.35 minutes to reach criterion. The extent to which vertical-horizontal division of labor increased inefficiency is thus clearly demonstrated. In addition, the added efficiency of the diagonal division of labor is apparent, however a Newman-Keuls test indicated that the difference between diagonal division and shared labor means did not reach significance at the .05 level. The significant Criteria effect reflected a general increase in the amount of time to fill in five words as the 25-word criterion was approached. The significant interaction between Criteria and Strategies indicates that this effect is true for the divided labor strategies but not for the shared labor strategy, which maintained a much more consistent pattern of performance across criteria.

The questionnaire items were combined to form "activity level," "interpersonal relations," and "task satisfaction" scores in a simple summation procedure described previously by Shiflett (1972). The analysis of variance of the activity level scores indicated that diagonal division of labor produced significantly lower activity ratings than did either the vertical-horizontal labor division or the shared labor condition ( $F = 13.05$ ,  $df = 2/24$ ,  $p < .001$ ). Vertical-horizontal labor division and shared labor produced virtually identical activity level ratings of 227.05 and 227.85 (vs. 186.40 for diagonal labor division). The substantial difference in task performance for these two conditions, coupled with their simi-

TABLE 2

SUMMARY OF ANALYSIS OF VARIANCE OF TIME-TO-CRITERION SCORES

Source	<i>df</i>	<i>MS</i>	<i>F</i>
Strategy (S)	2	24.01	
Error (between)	21	3.08	
Criteria (C)	4	6.03	3.03*
<i>S</i> × <i>C</i>	8	4.22	2.12*
Error (within)	56	1.99	

\* $p < .05$ .

\*\* $p < .01$ .

lar activity levels confirms the hypothesized deleterious effects of high task interdependence and low communicability.

The different labor strategies also significantly affected reported interpersonal relations ( $F = 13.45$ ,  $df = 2/24$ ,  $p < .001$ ), with shared labor producing the most positive ratings and diagonal division of labor producing the least positive ratings. This latter result, occurring among previously acquainted subjects, probably reflects the fact that there was very little interaction of any kind in the diagonal division of labor as a result of experimentally manipulated restrictions on communication. The analysis of variance of task satisfaction ratings produced no significant *F* ratios.

## DISCUSSION

The results have clearly demonstrated the superiority of the diagonal division of labor over the horizontal-vertical division with respect to both efficiency and effectiveness. The contention that the latter division introduced problems of high task interdependence with low communicability thus appears to be supported. These results also suggest that definite feedback regarding performance may improve substantially both efficiency and effectiveness. The same basic pattern of results reported by Shiflett (1972) was obtained for the shared labor and diagonal division of labor: The divided labor strategy was generally more efficient while the shared labor strategy was more effective. The hypothesis that divided labor would be equally effective on an easy task was not supported since shared labor

<sup>3</sup> As contrasted with a similar criterion of 45 words for the same type of analysis used in the previous study involving college students (Shiflett, 1972).

was more effective on both the easy and difficult tasks.

The superiority of the shared labor strategy may lie in the redundancy of the abilities of the two members that increased the probability that at least one member will have the correct solution, as suggested by Zajonc and Smoke (1959). However, on the more difficult task, the shared labor strategy, in which redundancy is maximized, failed to yield performance which significantly exceeded divided labor performance, where redundancy is effectively nil. This fact argues against the Zajonc and Smoke hypothesis and suggests that there may be a curvilinear relationship in which at the very easy and very difficult extremes redundancy is of little value, while at the intermediate levels redundancy is a major factor in increasing performance. At the easy extreme, overlapping ability is maximal but anyone working alone can do the same job as several persons working together while at the difficult extreme what is becoming highly redundant is not ability but the lack of it.

The diagonal division of labor can be viewed as a baseline strategy since that organization is essentially one of coacting individuals in which there is very little opportunity for the effects of either redundancy or interference to occur. The vertical-horizontal division can then be viewed as an example of the negative effects of a performance strategy, where the restriction on communication creates a situation in which an error by one partner either prevents solution or causes an error by the other member. The effect of this restriction seems likely to be very sensitive to task difficulty and member ability, since the probability of an error increases as task-relevant ability declines or difficulty increases. The shared labor strategy can also be viewed as containing an interfering or inefficient characteristic since the additional task of making a joint decision is required. But as the task progresses, the importance of this interference decreases relative to the redundancy advantages that accrue, so that in the later stages, and in terms of total performance, shared labor is the more effective strategy. Of course, the task must be such that redundancy is either helpful or necessary, otherwise neither effectiveness nor efficiency will benefit.

As shown in Figure 2, the vertical-horizontal division of labor performance curve never exceeds the shared labor or diagonal division of labor curve. In the original study, vertical-horizontal division of labor did exceed shared labor performance during the first few minutes and occupied an almost identical relationship relative to shared labor as does the present diagonal division of labor performance. It thus seems likely that had the diagonal division of labor been used in the original study, where average ability was much higher, the hypothesis regarding efficiency of dividing labor would have been even more clearly supported.

Inspection of Figure 1 indicates that, in terms of mean performance, the puzzles strongly differ in difficulty only during the first 6 to 8 minutes. After that time the difference in difficulty is small and nonsignificant. This same finding occurred in the original study, but the strong ceiling effect on performance that occurred there obscured this fact. Little, if any, ceiling effect operated in the present study, due primarily to the much lower ability level of the subjects (only two groups completed the easy puzzle). In general, there was little difference in difficulty between the two puzzles, as defined by word frequency, during the latter two thirds of the experimental period. The initially large differences in performance caused the tasks to remain significantly different in performance and, therefore, in perception of difficulty.

An additional problem with the definition of difficulty exists in the decline in performance over time which occurred on both the easy and the difficult task. This effect also occurred in the Shiflett (1972) study but was obscured by the fact that performance rapidly reached a maximal or near maximal level on the easy task due to the fact that so many groups nearly finished the task within 10 minutes. In the present study, performance again approached an asymptote, but the much lower level of performance here suggests that the ceiling effect is a reflection of lower ability levels rather than a task-imposed limitation. The substantially lower performance levels of the present groups as compared with the previous groups are consistent with the differences in pretest ability levels and suggests

that both tasks were, on the average, more difficult for the present subjects than for the original subjects. To the extent that performance level reflects task difficulty it can be argued that the difficulty of the task (filling in the remaining words) increases as the work proceeds. This effect probably reflects a tendency for subjects to fill in the easier words first and progress to the more difficult words within a puzzle.

A final and more general problem exists in the definition of task difficulty. The crossword puzzles were defined as if the property of task difficulty existed independently of the ability level of the individuals working on the puzzle. This is probably adequate in an ordinal sense since the difficult puzzle is relatively more difficult than the easy puzzle for almost all of the subjects used in these two studies, in terms of both performance and rating of difficulty. However, difficulty is also closely related to the relevant ability of the individual working on the task. Thus a task may be seen as difficult or even impossible to a person with little task-relevant ability but be seen as rather easy to a person with high ability. This same difference in perception can be expected to be reflected in actual task performance. Task difficulty, then, is relative to individual ability. Task difficulty can be defined relative to other tasks and relative to the ability of the persons performing the task. It has also been demonstrated that task difficulty may change in the course of working on the task. The failure to find that redundancy substantially aided group per-

formance in the difficult task but instead was more helpful on the easier task was perhaps the most surprising result of this study. In light of this finding, efforts to understand just how group organization and the distribution of resources within a group affect group performance and process may have to consider more carefully the effect of the interaction between task difficulty and member ability on those dependent variables.

#### REFERENCES

- DAVIS, J. H. *Group Performance*. Reading, Mass.: Addison-Wesley, 1969.
- GOLDMAN, M. A comparison of individual and group performance for varying combinations of initial ability. *Journal of Personality and Social Psychology*, 1965, 1, 210-216.
- LAUGHLIN, P. R., BRANCH, L. G., & JOHNSON, H. H. Individual versus triadic performance on a unidimensional complementary task as a function of initial ability level. *Journal of Personality and Social Psychology*, 1969, 12, 140-150.
- SHIFLETT, S. C. Group performance as a function of task difficulty and organizational interdependence. *Organizational Behavior and Human Performance*, 1972, 7, 442-456.
- STEINER, I. D. Models for inferring relationships between group size and potential group productivity. *Behavioral Science*, 1966, 11, 273-283.
- TAYLOR, D. W., & FAUST, W. L. Twenty questions: Efficiency in problem solving as a function of group size. *Journal of Experimental Psychology*, 1952, 44, 360-368.
- ZAJONC, R. D., & SMOKE, W. Redundancy in task assignments and group performance. *Psychometrika*, 1959, 24, 361-370.

(Received December 2, 1971)

## EXPERIMENTAL TEST OF THE VALENCE-INSTRUMENTALITY RELATIONSHIP IN JOB PERFORMANCE

ROBERT D. PRITCHARD<sup>1</sup> AND PHILIP J. DE LEO

*Purdue University*

Expectancy-valence models of work motivation postulate an interactive relationship between valence of outcomes and performance-outcome instrumentality. In order to test this postulate a laboratory simulation was created in which these two variables were experimentally manipulated. Valence of job outcomes was set at two levels, high and low, by establishing two different pay rates; performance-outcome instrumentality was determined by paying hourly (low instrumentality) or by the piece (high instrumentality). It was hypothesized that these variables would combine interactively to affect task performance and effort. While main effects for both performance-outcome instrumentality and valence of job outcomes were observed, the predicted interaction did not appear. One explanation of the data suggested that the typical conceptualization of valence as the importance an individual attaches to an outcome may be inappropriate.

Recently, a number of writers (Campbell, Dunnette, Lawler, & Weick, 1970; Galbraith & Cummings, 1967; Graen, 1969; Lawler, 1971; Porter & Lawler, 1968; Vroom, 1964) have applied expectancy-valence theories to the problem of work motivation. Such applications are special cases of the more general expectancy-valence approach that originated with Lewin (1938) and Tolman (1932). The amount of motivating force acting on a person to exert effort in work situations is generally viewed to be a function of three variables (*a*) valence of job outcomes—the attractiveness of the consequences of work performance for the individual, (*b*) performance-outcome instrumentality—the degree to which performance is related to obtaining each outcome, and (*c*) effort-performance expectancy—the perceived degree of relationship between effort and performance.

Expectancy-valence models postulate that these three variables combine multiplicatively. The valence for each job outcome is multiplied by the instrumentality of performance for attaining that outcome, and then these products are summed to obtain the valence attached to performance. Valence of performance is multiplied in turn by effort-performance expectancy, the result being a prediction

of force, or as it is usually operationalized, effort. These multiplicative relationships are critical to the expectancy-valence approach. They imply that if, for example, a person sees no relationship between his level of performance and the amount of money he earns, a potential pay raise will not affect his level of effort. In this case, while the valence of the pay raise itself may be very high, when this valence is multiplied by a performance-pay instrumentality of zero, the resulting product is zero. Thus, the outcome of receiving a pay raise serves to increase neither the overall value of high performance nor the force toward high effort.

If this relationship is additive instead of multiplicative, a completely different prediction is made. With an additive relationship the valued pay raise will increase effort no matter what the level of instrumentality or expectancy happens to be. The purpose of this study is to test the multiplicative relationship between valence of job outcomes and performance-outcome instrumentality.

Attempts have been made to test for the presence of this multiplicative relationship with correlational methods. The typical procedure has been to obtain ratings of the valence of job outcomes and the perceived degree of relationship between performance and obtaining the outcomes. One score is generated by multiplying the valence of each outcome by its instrumentality and adding

<sup>1</sup> Requests for reprints should be sent to Robert D. Pritchard, Assistant Professor, Department of Psychology, Purdue University, Lafayette, Indiana 47907.

the products. A second score is calculated by adding the valence of each outcome to its instrumentality and adding the sums. Each of these two scores is then correlated with performance and/or effort. If the multiplicative score produces higher correlations with effort and performance than does the additive score, support is claimed for the multiplicative relationship.

One example of this approach was reported by Hackman and Porter (1968). Using a sample of female telephone operators they found that the sum of the products of the instrumentalities times the valences correlated from .06 to .40 with a median of .27 with their various effort and performance criteria. An additive relationship calculated by summing valences and instrumentalities correlated from -.01 to .27 with a median of .17. (It should be noted, however, that their instrumentality measure was actually the relationship between effort and outcomes rather than between performance and outcomes.) Porter and Lawler (1968) and Pritchard and Sanders (1973) also found some support for the multiplicative relationship using this approach.

While such correlational tests are valuable, it seems necessary to use other approaches in testing this relationship. Specifically, it seems

necessary to test this relationship with experimental methods. This presents a problem, however, since to adequately test such a multiplicative relationship it is necessary to scale with precision the levels of the experimental treatments on both the valence of outcomes variable and the instrumentality variable. Only by specifying that the low-instrumentality condition is, for example, .2 and the high-instrumentality condition is .8, and by doing the same for valence, can a true test of the multiplicative relationship be made. It would be extremely difficult, for example, to accurately scale a piece-rate payment system (high instrumentality) as compared to an hourly payment system (low instrumentality).

The problem is greatly simplified, however, if one is satisfied to test for the presence of an interactive relationship between valence and instrumentality rather than for a true multiplicative relationship. Such an interactive prediction is presented in Figure 1.

As this figure indicates, it is predicted that increases in the valence of the outcome will result in a greater increase in effort and performance when instrumentality is high than when instrumentality is low.

The presence of such an interaction would support the presence of a multiplicative relationship. Furthermore, such an interactive relationship carries basically the same behavioral implication with it as does a truly multiplicative relationship. For example, the interactive relationship implies that a potential pay raise would have very little effect on effort if the relationship between performance and obtaining the pay raise was perceived to be low.

All that is really necessary to test for this interaction is to manipulate both the valence of job outcomes and the performance-outcome instrumentality so that two substantially different levels of each variable are present. In this study the job outcome of pay was selected, and the high-low instrumentality variable was operationalized as piece-rate payment and hourly payment, respectively. The valence variable was manipulated by offering different amounts of pay. It was assumed that the greater the pay, the greater would be the valence of pay. A 2 X 2 design

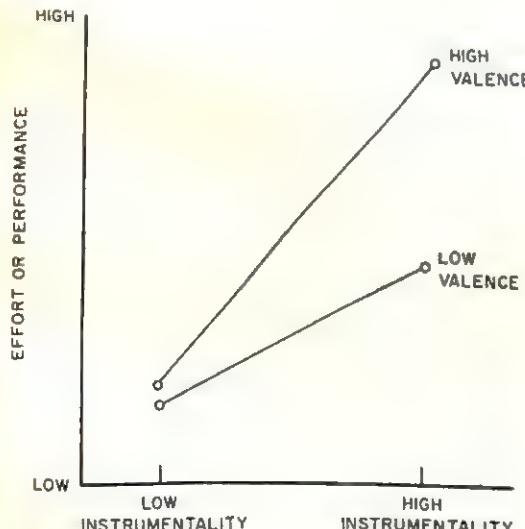


FIG. 1. Interaction predicted on the basis of multiplicative relationship between instrumentality and valence of outcome.

was thus generated with high and low instrumentality as one factor and high and low valence of pay as the other.

### METHOD

#### Subjects

Subjects were recruited by advertising for part-time clerical help in the local and campus newspapers. Those who served as subjects ( $N = 60$ ) were mainly college students (68%), predominately female (73%), and ranged in age from 16 to 27, with a median age of 21.

#### Task

The experimental situation was presented to the subjects as a real job. The experimenters<sup>2</sup> were older graduate students who looked and dressed like businessmen. Subjects had been told in the advertisement that they were being employed for a part-time job for one evening only. When they arrived for work, they were informed that the job they would perform simulated that of clerks in large mail order houses, and that they were being hired by the Occupational Research Center of Purdue for the purpose of determining cost and time data as well as information on people's reactions to this kind of job.

Most tasks permit output to vary simultaneously along two dimensions—quality and quantity. People may thus invest their motivational energies in either high quantity or high quality or some combination of both. Since it is desirable to be able to infer effort from output, an ideal task would vary on only one dimension. The task chosen for this study deliberately eliminated variation in quality so that only quantity would be reflected in the performance measure. The task employed was practically identical to that used in a study by Pritchard, Dunnette, and Jorgenson (1972). It involved transforming a catalog number by adding digits to it in accordance with a set formula, then looking up the transformed numbers in a 64-page "special sale" catalog to find a price corresponding to the translated catalog number. Each subject had been given a catalog and a set of worksheets. Each worksheet contained five untranslated numbers, the catalog page on which each of the five transformed numbers was to be found, and five pairs of prices. One of each pair of prices would be correct, would be found in the catalog on the given page, and was to be circled by the subject.

This task eliminates quality variation since it requires that each step of the task be done correctly before it is possible to finish one unit (five catalog numbers). If a subject incorrectly transforms a number, he will be unable to find that number in the catalog. Since there are five numbers and only five correct prices, if he circles an incorrect price, he will discover that he cannot find one of the other prices.

### Design

A  $2 \times 2$  design was employed with two levels of instrumentality (piece-rate payment and hourly payment) and two levels of valence of pay. Subjects in the hourly, low valence of pay condition were paid \$1.75 per hour and subjects in the high valence of pay condition, \$2.50. It was felt that for the piece-rate condition the pay per piece should be set in such a way that if a subject in the piece-rate condition performed at the same level as the mean of the hourly group, he should receive the same pay as the hourly group. This ensured that when performance was equal the level of rewards for the piece-rate groups would be equal to the hourly groups. The rate paid the two hourly groups (\$1.75 and \$2.50) was thus divided by the group's mean hourly performance, and rounding to the nearest whole cent, this resulted in \$.07 per piece for the low-valence condition and \$.10 for the high-valence condition.

### Procedure

Subjects had been told in the advertisement to report to a room and building on campus. When they arrived, they were greeted by two experimenters. When 30 subjects had arrived, they were split randomly—half remaining where they were and half accompanying one experimenter to an identical classroom. The two hourly groups ( $n = 15$  each) were run on one evening and the two piece-rate groups ( $n = 15$  each) were run in the evening 1 week later. Where more than 30 subjects reported, those reporting last were told that all positions had been filled. They had been warned of this possibility in the advertisement.

After a demonstration of the task by the experimenter the subjects were given a short practice period and then they were given an actual job sample.

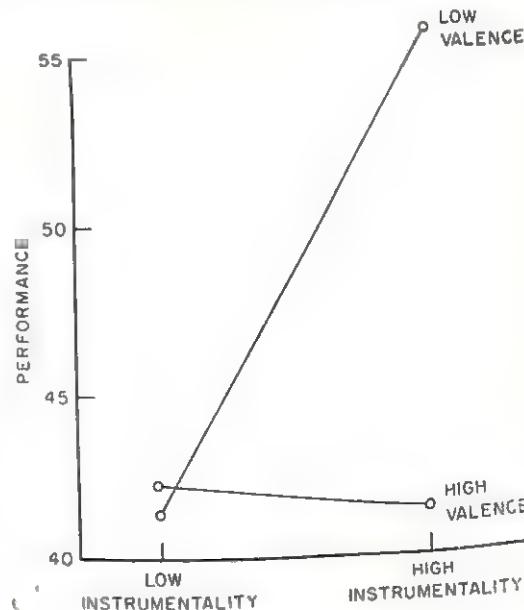


FIG. 2. Task performance, by treatment.

<sup>2</sup> The authors would like to express their appreciation to Thomas L. Hozman and Robert R. Wood, who served as experimenters.

They were urged to do their best, since "successful performance on this task would be a necessary condition for employment." The number of units completed in this 15-minute period was actually to serve as an ability measure. The purpose of inducing this selection set was to ensure that subjects would exert maximum motivation for the ability pretest. All subjects were indeed "hired."

At the completion of this pretest, the experimental induction was given, that is, the rate of pay was announced and the type of pay system was explained. The subjects worked steadily at the task for 90 minutes. At that point they were given a questionnaire that measured, among other things, the effectiveness of the manipulations. Subjects were then paid and dismissed. The subjects were not suspicious about the nature of the "job," and the only questions asked of the experimenters dealt with possibilities of future employment. Since real deception was not involved, the subjects were not debriefed.

## RESULTS

### *Checks on the Manipulations*

Since performance-pay instrumentality and valence of the outcome of money were manipulated, an attempt was made to check the effectiveness of these manipulations by a postexperimental questionnaire. To check the instrumentality manipulation subjects were asked: "If you were to increase your performance on this job (finish more blocks), what are the chances in 10 that you will make more money?"

The mean response to this question for subjects in the high-instrumentality (piece-rate) pay condition was 7.70 and for subjects in the low-instrumentality condition it was 2.85; this difference was highly significant ( $p < .001$ ). These data suggest that the instrumentality manipulation was highly effective. One might argue that since the hourly condition was actually seen to have an instrumentality greater than zero, the theory would predict that there should be a small difference between the high- and low-valence conditions in the hourly pay system. It seems doubtful, however, that such a low instrumentality (2.85) would result in measurably different effects on performance.

In order to assess whether subjects saw a significant difference in the two pay rates, subjects were asked to rate how attractive various pay rates for this job would be to them. Subjects rated each of seven hourly rates (\$1.50, \$1.75, \$2.00, \$2.25, \$2.50, \$2.75,

and \$3.00) on a 9-point Likert scale ranging from "unattractive" to "extremely attractive." To test whether \$2.50 was seen as more attractive than \$1.75 the means of the ratings given by all subjects on those two pay levels were compared. A  $t$  test for dependent measures indicated that the \$2.50 pay rate was seen as significantly ( $p < .01$ ) more attractive ( $\bar{X} = 5.53$ ) than the \$1.75 ( $\bar{X} = 2.37$ ). These data suggest that the \$1.75 was seen as lower in valence than the \$2.50 pay rate. The attractiveness of the two piece rates was not actually assessed since, when performance was equivalent, those piece rates corresponded to the two hourly pay rates. The assumption could thus be readily made that the two piece rates were also seen as different in attractiveness.

### *Tests of the Hypotheses*

The expectancy-valence model predicts that (a) there should be no difference in performance between the two levels of pay for subjects in the low-instrumentality (hourly) condition; (b) performance should be higher for both pay-level conditions in the high-instrumentality (piece-rate) condition than in the low-instrumentality condition; and (c) within the high-instrumentality condition the high-pay group should outperform the low-pay group. These predictions are represented graphically in Figure 1.

Figure 2 presents the actual performance data for the four conditions. The results do not support the predictions. While the analysis of variance indicated there was a main effect due to instrumentality ( $F = 7.61, df = 1/56, p < .01$ ), the cell means show that this effect is due to subjects in the low-valence-high-instrumentality condition strongly outperforming all other subjects. The high-valence-high-instrumentality subjects did not demonstrate higher performance than the low-instrumentality-high-valence subjects as had been predicted.

The overall test of the interactive relationship is the interaction between valence and instrumentality. Although this two-way interaction was significant ( $F = 9.24, df = 1/56, p < .005$ ), the obtained pattern of means clearly does not conform to the predicted pattern.

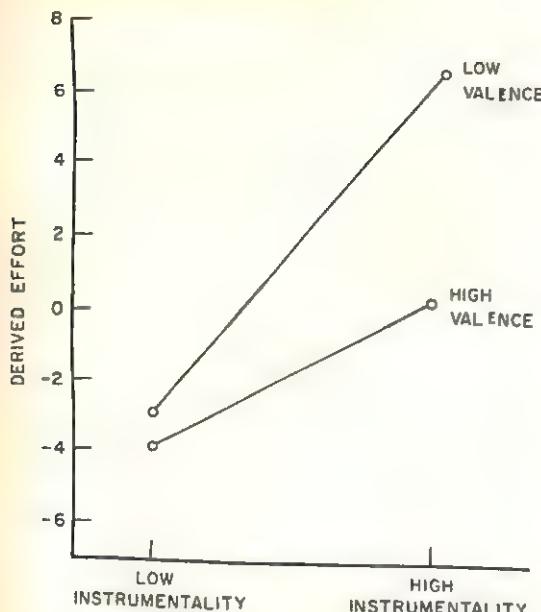


FIG. 3. Derived effort, by treatment.

While actual performance is an important dependent variable for expectancy-valence models, the models actually attempt to predict effort. Consequently, an attempt was made to obtain a measure of effort on the task. If one assumes that performance on a task such as this is largely a function of ability and motivation (or effort), then partialing out ability in some fashion should yield a measure of effort.

After being familiarized with the task, the subjects had 15 minutes to complete as much of the task as possible. They were told that this "test" would determine whether or not they would be hired for the job. In fact, the pretest was designed as a measure of the subjects' ability to perform the task.

It was assumed that performance on this pretest was a measure of ability, and these data were used to produce derived effort scores. This was accomplished by generating a regression equation which predicted performance on the actual 90-minute task from the ability pretest score.<sup>3</sup>

Using the regression equation generated in this fashion, a predicted performance score was calculated for each subject from knowledge of his ability pretest score. Finally, each

subject's predicted score was subtracted from his actual score. This deviation score was then considered to be a measure reflecting the level of effort a subject expended on the task relative to other subjects in the experiment. For example, if Subjects A and B had the same level of ability and Subject A outperformed Subject B, one could assume that Subject A had exerted a higher level of effort than Subject B. Since both subjects would receive the same predicted score due to their equal ability, the actual score minus the predicted score would be higher (more positive or less negative) for Subject A, thus reflecting his greater effort.

These derived effort scores were also analyzed with a two-way analysis of variance. The cell means are shown in Figure 3. These data support the prediction of high effort under the high-instrumentality condition ( $F = 15.30, df = 1/56, p < .001$ ). The effects of valence, however, were significantly ( $p < .05$ ) opposite those predicted. While the cell means show an interaction, this effect achieved a  $p$  value of only .08.

These results clearly do not support the hypothesis. The predicted interaction did not appear, and the low-valence subjects either equaled or exhibited higher performance and effort than high-valence subjects. However, there was positive evidence for the effects of instrumentality.

#### DISCUSSION

Clearly the data do not provide immediate support for the expectancy-valence model. Determining what the data do support, however, is a difficult matter.

There are several possible interpretations: The first possibility is that the manipulation of valence was inadequate. However, several factors argue against this. Since the manipulation check was highly significant, it indicates that these subjects perceived \$1.75 per hour as less attractive than \$2.50. Furthermore, the \$2.50 rate is 40% higher than the \$1.75, and if differences could not be detected with this strong a manipulation, the model is of questionable use in explaining behavior. These factors tend to rule out the possibility of an inadequate valence manipulation.

<sup>3</sup> The correlation between ability pretest and task performance was .64.

A second explanation is the possibility that effects due to feelings of inequity (Adams, 1965) acting in conjunction with expectancy-valence effects produced the findings. One of the essential elements in equity theory, however, is the comparison object. In this case it is doubtful that subjects could have known what subjects in other treatments were being paid. The high- and low-valence conditions of each instrumentality level were run at the same time, and the two instrumentality levels were run 1 week apart. While it is possible that some subjects had heard about the pay levels under the hourly conditions, this was probably rare if it occurred at all.

Another problem with an equity interpretation of these data is that, at least in the high-instrumentality conditions, the nature of the task made underpayment inequity reduction very difficult. Since high valence showed lower performance and effort than low valence, these high-instrumentality groups are the central source of the data disconfirming the expectancy-valence predictions. Yet to posit that the low-valence group felt underpaid and thus performed highly on the piece-rate pay system to reduce feelings of underpayment is not likely. The typical finding in the equity literature (Lawler, 1968; Pritchard, 1969) is that subjects do increase quantity of performance under piece-rate underpayment, but with lower quality of performance. Thus, they are seen as doing poorer quality work very rapidly so as to reduce feelings of inequity. In this study, however, the task was structured so that it was impossible to do lower quality work since to finish one unit of the task it had to be done correctly. Because quality could not be lowered, increased performance could not reduce feelings of inequity for the low-valence subjects. The high effort and performance of the low-valence-high-instrumentality subjects was, therefore, probably not due to equity effects.

A third possible explanation for the findings reported here is that the valence component of expectancy-valence models is not a critical component and, in fact, does not add to prediction. This possibility is supported by some correlational studies testing expectancy-valence models (e.g., Gavin, 1970; Jorgenson,

1970). However, other evidence does indicate that job performance and effort are related to the valence of job outcomes (Lawler & Porter, 1967; Porter & Lawler, 1968; Pritchard & Sanders, 1973). One problem with this interpretation is that the differences between low- and high-valence in the high-instrumentality condition were actually significant, but in the opposite direction. If valence of outcomes were not an important aspect of the model, a difference should not have emerged. A more complex explanation is clearly necessary.

One curious finding, which eventually suggested a fourth explanation for the data, was that in the piece-rate condition both the low- and high-valence groups actually earned almost the same amount of money. The difference between the groups in the total amount earned for the entire experimental period was about \$.20. This would imply that the low-valence group was willing to work harder to earn the same amount of money. It was almost as if the two groups had an equal need for earning that amount of money and strove to do so even though it required greater effort for the low-valence group.

In more general terms this argument says that the level of need a person has for an outcome must also be considered as a determinant of his perceived valence for that outcome. This is different from the way valence is normally conceptualized. In most studies (e.g., Porter & Lawler, 1968; Pritchard & Sanders, 1973) subjects are asked to indicate the *importance* of various outcomes. We are arguing that, in addition to importance, the valence of a particular job outcome is also determined by the *level of need* for that outcome. For example, a person may feel recognition is more important than salary and indicate such on a questionnaire. However, if he feels he is receiving enough recognition on his job at a given point in time, he may need or want a pay raise more than increased recognition.

Other research supports this line of reasoning. Lawler and O'Gara (1967) found that subjects' performance on a piece-rate pay system was positively related to their need for money. Andrews (1967) reported that previous wage history was positively correlated with piece-rate performance. These

findings also support our line of reasoning to the extent that previous wages are related to need for/or attractiveness of money. Finally, Lawler (1971) suggested that as more money is earned, the need for money decreases. Thus, as subjects earned more and more money in the high-piece-rate condition, their need for money could have decreased. This decrease would not be as large for subjects in the low-piece-rate condition since they were earning less money.

This argument suggests that to measure the valence of the outcomes component in expectancy-valence models one should measure the *need* a subject has for the outcome, and perhaps better yet, the need he has for a specific level of that outcome. Instead of, or in addition to, asking how important a salary raise is to the individual, one should ask how badly does he want a salary increase at this time, or how badly does he want a salary raise of \$75 per month.

This line of argument would imply that in the present study the need for earning money was possibly the same for both high- and low-valence groups, and that they performed at a level necessary to satisfy this need. This explanation of the findings would further imply that if one could group subjects on the basis of their level of need for earning money or level of need for earning specific amounts of money, one would have a better operationalization of valence than the one used in the study.

While the research presented here cannot be said to have supported the interactive or multiplicative relationship between instrumentality and valence of outcomes, it at least suggests that the conceptualization of valence used in previous research may be incomplete and that a more appropriate measure should include the idea of need for the outcome in addition to the idea of importance of the outcome.

#### REFERENCES

- ADAMS, J. S. Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 2. New York: Academic Press, 1965.
- ANDREWS, I. R. Wage inequity and job performance: An experimental study. *Journal of Applied Psychology*, 1967, 51, 39-45.
- CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial behavior, performance and effectiveness*. New York: McGraw-Hill, 1970.
- GALBRAITH, J. R., & CUMMINGS, L. L. An empirical investigation of the motivational determinants of task performance: Interactive effects between instrumentality-valence and motivation-ability. *Organizational Behavior and Human Performance*, 1967, 2, 237-257.
- GAVIN, J. F. *Ability, effort, and role perceptions as antecedents of job performance*. (Experimental Publication System Ms. No. 190A) Washington, D.C.: American Psychological Association, 1970.
- GRAEN, G. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology*, 1969, 53(No. 2, Pt. 2).
- HACKMAN, J. R., & PORTER, L. W. Expectancy theory predictions of work behavior. *Organizational Behavior and Human Performance*, 1968, 2, 417-426.
- JORGENSEN, D. O. An experimental and correlational analysis of some expectancy-valence models of motivation. Unpublished doctoral dissertation, University of Minnesota, 1970.
- LAWLER, E. E. Equity theory as a prediction of productivity and work quality. *Psychological Bulletin*, 1968, 70, 596-610.
- LAWLER, E. E. *Pay and organizational effectiveness: A psychological view*. New York: McGraw-Hill, 1971.
- LAWLER, E. E., & O'GARA, P. W. Effects of inequity produced by underpayment on work output, work quality, and attitudes toward the work. *Journal of Applied Psychology*, 1967, 51, 403-410.
- LAWLER, E. E., & PORTER, L. W. Antecedent attitudes of effective managerial performance. *Organizational Behavior and Human Performance*, 1967, 2, 122-142.
- LEWIN, K. *The conceptual representation and the measurement of psychological forces*. Durham: Duke University Press, 1938.
- PORTER, L. W., & LAWLER, E. E. *Managerial attitude and performance*. Homewood, Ill.: Dorsey Irwin, 1968.
- PRITCHARD, R. D. Equity theory: A review and critique. *Organizational Behavior and Human Performance*, 1969, 4, 176-211.
- PRITCHARD, R. D., DUNNETTE, M. D., & JORGENSEN, D. O. The effects of perceptions of equity and inequity on worker performance and satisfaction. *Journal of Applied Psychology*, 1972, 56, 75-94.
- PRITCHARD, R. D., & SANDERS, M. S. The influence of valence, instrumentality, and expectancy on effort and performance. *Journal of Applied Psychology*, 1973, 57, 55-60.
- TOLMAN, E. C. *Purposive behavior in animals and men*. New York: Appleton-Century, 1932.
- VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.

(Received January 3, 1972)

# EFFECTS OF THE MANIPULATION OF A PERFORMANCE-REWARD CONTINGENCY ON BEHAVIOR IN A SIMULATED WORK SETTING<sup>1</sup>

DALE O. JORGENSEN<sup>2</sup> AND MARVIN D. DUNNETTE

*University of Minnesota*

ROBERT D. PRITCHARD

*Purdue University*

Research reported here was aimed at testing predictions derived from several Expectancy  $\times$  Value theories of motivation. Experimental manipulation of a performance-reward contingency was carried out on a sample of 256 male college students who were hired to work 6 consecutive days under simulated work conditions, 3 days under a high performance-reward contingency condition and 3 days under a low contingency condition. This manipulation was examined for its effects on the subject's perceived effort-pay probability, perceived effort, performance, and valence of pay. As predicted, this manipulation had a significant effect on effort-pay probability and performance, but failed to produce the predicted differences in perceived effort. The effects on valence were mixed.

The beliefs that an individual has about the consequences of a certain act or set of acts has recently attracted considerable attention from investigators of work behavior (Campbell, Dunnette, Lawler, & Weick, 1970; Graen, 1969; Porter & Lawler, 1968; Vroom, 1964). It is within the framework of the "Expectancy  $\times$  Value" theory that the impact of such beliefs has been most clearly explicated. Most prominent in the early development of the Expectancy  $\times$  Value theory were Tolman (1932) and Lewin (1938). Basic to both efforts was the notion that animals and humans have cognitive expectancies or anticipations about the outcome(s) of a certain act or set of acts that they might undertake. The other major variable in these models, value, implies that, in addition to subjective beliefs about the consequences of an act, organisms also have valuations of these consequences that may vary from strongly positive to strongly negative.

The motivation or force to undertake a particular act is then seen as a function of some combination of these two sets of variables. This implies that, with everything else constant, an individual who has a high level of expectancy for one outcome should exert more effort or would be subject to greater force to engage in a certain action than an individual with a lower level of expectancy for that same outcome. It was this type of motivational model, as it applies to behavior in work settings, that this investigation was designed to test.

The major hypotheses tested in this investigation were those based on an experimental manipulation of a performance-outcome contingency. As indicated, one of the basic postulates of all Expectancy  $\times$  Value theories is that individuals who have high expectancies, however defined, will behave differently than individuals who have low expectancies, provided that other variables (e.g., valence) remain constant. It follows that, if an objective performance-outcome contingency is manipulated by creating one condition in which there is a high performance-outcome contingency and another condition in which this contingency is lower, there ought to be differences in the strength or level of the expectancies and consequently, in the level of effort and performance between individuals in these two conditions. With pay as the outcome, the follow-

<sup>1</sup> Support for conducting this research came from National Science Foundation Grant GS 1862. The authors wish to express thanks to Robert O. Opsahl, who originally conceived the idea of conducting this experiment and who aided us considerably during our early planning.

<sup>2</sup> Requests for reprints should be sent to Dale O. Jorgenson, Department of Psychology, California State University, Long Beach, Long Beach, California 90840.

ing hypotheses flow from the basic assumption stated above:

Hypothesis 1a. With abilities and role perceptions constant, individuals under high performance-outcome contingency conditions have higher perceived effort-pay probabilities than individuals under low contingency conditions.

Hypothesis 1b. With the valence of pay and all other outcomes equal, individuals who are under high performance-outcome contingency conditions exert higher effort in performing the task than individuals who are under low contingency conditions.

The Porter and Lawler (1968) model also postulates that effort, in combination with certain other variables, influences performance (quality and quantity). Consequently, the following should also be true:

Hypothesis 1c. With the valences of all outcomes constant, individuals who are under high performance-outcome contingency conditions have a higher level of quantitative job performance than the individuals who are under low contingency conditions.

In line with these hypotheses, a shift in the actual performance-outcome contingencies from high to low or from low to high should lead to some changes in effort, performance, and the perceived contingencies.

Hypothesis 1d. When individuals under high performance-outcome contingency conditions are shifted to low contingency conditions, their perceived effort-pay contingencies and, therefore, their effort and performance levels will be lower after the shift than before, and also decrease over time, provided that valences remain constant; when individuals under low contingency conditions are shifted to high contingency conditions, their perceived effort-pay contingencies and, therefore, their levels of effort and performance will be higher after the shift than before and also increase over time, provided that valences remain constant.

The predictions of changes in the dependent variables over time after the shift are based on the feedback loops hypothesized in the Porter and Lawler (1968) model. The idea is that over time, perceptions of expectancy should come to approximate the objective situation and that

these changes in expectancy will have an influence on effort and performance.

This completes the set of hypotheses that follow from the experimental manipulation of the objective performance-pay contingencies. As such, they are the most crucial predictions of this investigation, mainly because of their implications in testing for the presence of causality between an individual's behavior and his perceptions of the relationship between his actions and some specified set of outcomes.

With the data collected in this investigation, it was also possible to test correlational hypotheses. Most of these hypotheses were based on the Porter and Lawler (1968) model. As viewed in all  $\text{Expectancy} \times \text{Value}$  formulations, the value or valence of an outcome is as important as expectancy level in determining behavior. In the Porter and Lawler model, as in other formulations, the valence variable is hypothesized to interact multiplicatively with expectancy or effort-reward probability. The following hypotheses were suggested by this basic postulate of  $\text{Expectancy} \times \text{Value}$  theory.

Hypothesis 2a. With valence constant, the higher the perceived effort-reward probability, the greater the effort expended by an individual in performing his job and the greater his job performance.

A similar prediction can be made for the valence variable.

Hypothesis 2b. With expectancy or effort-reward probability constant, the higher the valence of an outcome, the greater the effort expended by an individual in performing his job and the greater his performance.

## METHOD

### Subjects

The subjects were 256 undergraduate college males who answered advertisements for part-time employment with a temporary manpower firm over their spring vacation. Because of the potential risk of interaction between subjects in different treatments and because of the need for large numbers of subjects, several experimental sites were set up in various cities in a midwestern state. An attempt was made to recruit subjects from schools with fairly similar student body populations. In this case, it was limited to students enrolled in five state colleges. All applicants who appeared for the first day were admitted except in the case of three or four applicants at one site who were turned away because there were already enough subjects present.

### Task

Two criteria were employed in selecting and pretesting the task. The first was that it be realistic enough so that the job would be perceived as a real job. It was hoped that a realistic task would eliminate the effects of experimental demand on the subjects (see Orne, 1962). Secondly, it was felt that the task itself should have little or no quality variation. At the very least it was hoped that it took as much time and effort to do it incorrectly as correctly. The reasoning behind this second criterion was that the results of studies in which both quality and quantity were dependent variables are quite confusing.

The final form of the task required the subjects to find the prices of selected items in a mail order catalogue and then to indicate the correct price of each item on a standardized work sheet.

The subjects were instructed to transform or "decode" an item identification number using a simple addition rule, turn to the page in the catalogue listed for that item, find the item and its sale price, then circle the correct price from the five pairs of prices listed beneath each block of items. Finally, he was to indicate the item to which the circled price corresponded by writing a 1, 2, 3, 4, or 5 next to the circled price. Once all five items in a block had been completed the subject was to proceed to the next block of five and repeat the above procedure.

This kind of task eliminated quality variance. This was due to the fact that a subject was forced to correctly transform the identification number if he was to find the correct price in the catalogue. Once the item was found in the catalogue, he merely circled the one correct price from the pair printed on the sheet and indicated the item to which it corresponded. No errors were found in the pretest data nor in spot checks made during the study.

*Work Setting.* Every attempt was made to make the work situation seem realistic. To begin with, subjects were told that the company was a newly established manpower overload firm which mainly contracted for clerical work with companies who did not have the facilities or the time to complete this work. The company name was displayed on all advertisements for the job, on the checks used to pay the subjects, and on the printed time cards that were completed by the subjects every day. A rationale was given for the item identification decoding. This was, that the merchandise was sale merchandise; hence, different identification numbers were used than in the regular nonsale catalogues. It was also explained that while most of the work was usually done by a computer, the particular department store involved was currently experiencing difficulties with its computer, and because of a massive backlog of unprocessed orders had hired the manpower overload firm to do this work. Furthermore, actual sale catalogues from a well-known national retail chain were used, and the sale termination date given on the cover of the catalogue coincided with the time of the study. The task material consisted of actual computer from the company's computer output. In addition, the experimenters were carefully trained and rehearsed and were also given a very detailed 26-page manual of instructions that outlined

procedures to be followed closely and included a set of possible questions subjects might ask. Each experimenter was thoroughly familiar with this manual. It was hoped that by employing procedures such as the above many of the features of the usual employment setting would be retained while at the same time, loss of experimental control, so characteristic of many field experiments (see Weick, 1967) would be minimized.

### Procedure and Design

When subjects reported to the site they were met by a male and female experimenter who gave them a one-page description of the "company" and an application blank to complete. When all subjects had finished the application blank, the male experimenter who played the role of "supervisor" introduced himself and his "secretary," a role played by the female experimenter, by giving some background information about the firm, and then explained what the subjects would be doing for the remainder of that day and on the 6 subsequent days. The subjects were told that while the main contract was with the chain department store, another contract had been arranged with Science Research Associates (SRA) to pretest some tests of reactions to routine work. This experimental set was given to make the multitude of questionnaires which were given seem reasonable and also make the job more realistic.

When this set of questionnaires had been completed, subjects were given the three Short Employment Tests (SET): numerical, clerical, and vocabulary (Bennett & Gelinck, 1956). When the tests had been completed, the female experimenter left the room to supposedly score the SET tests while the male experimenter explained the mechanics of the catalogue task. The subjects then completed one page of "practice material," which allowed the experimenter to insure that all subjects were doing the task properly. When all had finished the single sheet of four blocks, the subjects took a short break.

After the break, the rationale for the catalogue task was given along with a 1-hour pretest on the task. Both the SET and this 1-hour task pretest were to be measures of ability but not actual selection instruments. However, to insure high motivation as well as to provide explanation for their administration subjects were told that they would be used for selection purposes. In fact, all subjects were hired.

The method of payment advertised as \$2.00 per hour for groups in the hourly conditions and "from \$1.60 to \$2.40 per hour depending on what you do" for the incentive pay groups, was then explained to the subjects. Regardless of the treatment condition, the subjects were paid for the two contracts separately. Pay for work done on the catalogue job each day was given on the subsequent morning while payment for the first day's introduction and testing as well as all the money for the SRA contract (all of this at \$2.00 per hour) was paid at the end of the last day. There were several reasons for this complex method of payment. The subjects were paid every day to provide feedback about the performance-pay contingency in each of the expectancy treatment conditions. Also, when the switch in pay system occurred, any change in amount of money earned would be

very obvious. In addition, the first day's pay was held until the end to encourage subjects to keep coming back.

When the subjects arrived for work 2 days later (the original data collection took place on a Saturday), experimenter briefly reviewed the directions for the task and explained that their finished work would be picked up each hour since experimenter and the secretary had "to process it before it is sent out."

At this point, the expectancy manipulations were given. The expectancy manipulation was carried out by varying the actual contingencies between one single outcome, amount of pay, and overt performance as measured by the number of "blocks" of five items completed per unit of time. In this case, two different pay-performance contingencies were arranged. The *low expectancy* or *low performance-pay contingency* condition, was created by simply paying subjects on an hourly basis. The *high expectancy* or *high performance-pay contingency* condition was created by paying subjects on a modified piece rate or incentive system.

When subjects worked under the hourly condition, they received a straight \$2.00 per hour. The incentive system was more complex. If a subject completed between 16 and 22 five-item blocks in a given hour, he would receive \$1.60 for that hour; if he completed between 23 and 29 blocks, he received \$2.00; and finally, if he completed 30 or above, he was paid \$2.40. However, if he produced below 16, he received only 8¢ per block. The latter system was irrelevant since pretest data indicated that all subjects tested could produce 16 blocks or more after practice.

The pay rates used were based on several considerations. The difference between the pay of the three performance intervals had to be large enough in order to make it worthwhile to strive for a higher interval, yet not so large as to make the false pay rates used in the equity manipulation (which were higher or lower than the actual rates) seem unbelievably high or low. It was felt that anything below \$1.35 or above \$3.00 per hour would not seem reasonable to the subjects. In fact, the actual hourly and the middle interval rate of \$2.00 was based on what college students in the pilot studies felt was a fair rate of pay for the task. The pilot studies also indicated that a difference of 30¢ to 40¢ between the intervals was more than sufficient to motivate subjects to try for the higher intervals since the subjects in these pilot studies did indeed strive for the higher intervals. Thus, using the base of \$2.00, the actual rates were set at \$1.60, \$2.00, and \$2.40 for the interval pay system, and \$2.00 for the hourly system. The performance intervals were also based on the pilot studies and were set so as to maximize the chances that an equal number of subjects would fall into each interval.

### *Low Expectancy*

One hundred and five subjects in three equity conditions, underpayment ( $n = 22$ ), equity ( $n = 58$ ), and overpayment ( $n = 25$ ), worked the first 3 days under this low expectancy condition. They were told that their pay, as advertised, would be \$2.00 per hour, "the rate we have been paying college students for this work." They then started working on the catalogue task.

### *High Expectancy*

Once again, subjects in three equity treatments ( $n = 25, 48, 18$ , respectively) worked the first 3 days under this condition. They were told that the rates listed earlier were the rates that the company had been paying college students for this work. After the payment manipulation had been given, subjects began work on the catalogue task. They worked for a total of 4 hours on each of 6 days, with their output being collected at the end of each hour. At the end of the fourth hour each day, they were given the set of questionnaires that were supposedly part of the SRA contract.

The six payment conditions that resulted from crossing the two expectancy treatments with the three equity treatments were employed for the first 3 actual working days on the job. The very first day, of course, had consisted of the ability tests, learning the job, and the 1-hour work sample. After 3 days of actual work on the job, all groups were shifted to the opposite expectancy condition for the final 3 days. All groups who started in the hourly condition were shifted to the interval condition, and all groups starting on the interval condition were shifted to hourly. The intervals, the amount of pay for each interval, and the hourly rate were the same as those used in the first 3 days.

All shifts were accomplished by simply telling the subjects on the morning of the fourth day that the main office had decided to change the pay system. The new system was then explained.

*Equity Manipulations.* As discussed previously, the two expectancy conditions were crossed with three equity conditions. These conditions will not be discussed here since they have been discussed previously (Pritchard, Dunnette, & Jorgenson, 1972) and since the focus of this report deals with the expectancy data.

### *Measurement Instruments*

To test the hypotheses, several measurement instruments were developed. The variables which were measured as part of the investigation were as follows: (a) effort-pay probability, (b) performance-pay probability, (c) valence of job outcomes, and (d) effort.

### *Effort- and Performance-Pay Probability*

A questionnaire was developed to measure both the subject's subjective probability that increased effort would lead to a larger amount of pay as well as the probability that increased performance would lead to greater pay. This was done by having the subject indicate a number from 0 to 100 that would reflect the extent to which he felt that amount of effort and amount of performance leads to an increased amount of pay. For example:

The chances are \_\_\_\_ in 100 that a person who puts in a lot of effort will make more money than a person who only puts in a little effort.

In the case of performance expectancy, the corresponding statement reads:

The chances are \_\_\_\_ in 100 that a person who finishes a lot of work on this job will make more

money than a person who finishes a small amount of work.

*Valence of Job Outcomes.* Measures of the valence of pay as well as 12 other job outcomes were taken. These additional outcomes included making use of abilities, feeling of accomplishment, being busy all the time, fairness of company policy, friendliness of co-workers, freedom to try own ideas, opportunity to work alone, recognition for work done, opportunity to make decisions, having a boss who backs up his workers, doing something different every day, and good working conditions.

Measures of valence were obtained by means of a modified Q-sort method (Stephenson, 1953). The task for the subject was to place each of 13 outcomes, (pay, recognition, security, etc.) in one of five *a priori* categories based on the importance of each outcome to the subject. The five categories were "Not a necessary part of a job"; "Okay to have on a job but not really important"; "Rather desirable in a job"; "Highly desirable in a job"; "Absolutely necessary in a job."

#### Effort

The subject's perception of the extent to which each of the nine job inputs, including effort, was present on this job was measured by means of a multiple rank order paired comparison method (Gulliksen & Tucker, 1961). This method permits a choice of response formats based on the number of statements per group to be rank ordered. For each of these formats there is a specific number of statement combinations which pairs each statement with each other statement once and only once. The task for the subject was to rank order nine triads of the inputs in terms of "how much you feel you bring the characteristic to the job." The perceived level of effort was measured by the number of times each input was ranked over the remaining eight inputs, adjusted for the neutral point.

#### RESULTS

All of the experimental hypotheses tested herein required that the valence of money be equal under high and low expectancy conditions. Furthermore, none of the expectancy value models predicted any differences in valence due to differences in expectancy condition with the exception of the Atkinson (1958) model. The lack of a significant difference in the valence of money due to expectancy condition supports this requirement ( $F = 1.77$ ,  $ns$ ,  $df = 1,139$ ,  $\omega^2 = .001$ ).<sup>3</sup> However, examination of the means in Table 1 will reveal that significant differences between high and low ex-

<sup>3</sup> The  $\omega^2$  (omega squared) statistic is interpreted as an estimate of the proportion of variance in the dependent variable attributable to the variation in the independent variable. Details of the calculation of this measure are available from the first author.

TABLE 1  
MEAN VALENCE OF MONEY IN HIGH AND LOW EXPECTANCY CONDITIONS FOR GROUPS SHIFTING FROM HIGH TO LOW AND FROM LOW TO HIGH EXPECTANCY CONDITIONS

Order	Treatment condition	
	High expectancy (HE)	Low expectancy (LE)
High—Low (Order 1)	11.15	11.66
Low—High (Order 2)	11.30	10.60

Note. Individual comparisons are as follows: HE Order 1 vs. LE Order 1,  $F = 4.41$  ( $p < .05$ ); HE Order 2 vs. LE Order 2,  $F = 8.30$  ( $p < .01$ ).

pectancy conditions within order of treatment were masked when the means for the two separate orders were combined.<sup>4</sup> These differences were responsible for the significant Order  $\times$  Expectancy interaction ( $F = 11.96$ ,  $df = 1,139$ ,  $p < .01$ ). For the group which shifted from high to low expectancy, there was a significant increase in the valence of money after the shift to low expectancy. There was also a significant increase in this variable after the shift in the low  $\rightarrow$  high group. The difference lies in the fact that in the first group the mean was greater under low expectancy than under high, whereas in the second group it was greater under the high condition than under the low. Consequently, there was a general trend for subjects to perceive money as being more important over time relative to other outcomes regardless of the contingency condition under which they performed. This is reflected in the significant increase in the high expectancy group over time shown in Table 2. Thus, even though the requirement of no difference in valence is supported in substance, there is some evidence that the valence of pay does change over time.

In the hypotheses regarding performance as a dependent variable, it has been assumed that variables other than the valence of pay are equal in high and low expectancy conditions. If these assumptions are not met, at least some behavior differences might be partially attributable to differences in these other vari-

<sup>4</sup> The formula for calculating the  $F$  tests on all the *a priori* comparisons was derived and is available from the first author.

TABLE 2

MEAN VALENCE OF MONEY IN HIGH AND LOW EXPECTANCY CONDITIONS ACROSS THE TWO DAYS

Day	Treatment condition	
	High expectancy (HE)	Low expectancy (LE)
1	10.76	11.06
2	11.70	11.18
X	11.14	11.04

Note. Individual comparisons are as follows: HE Day 1 vs. HE Day 2,  $F = 19.38$  ( $p < .01$ ); LE Day 1 vs. LE Day 2,  $F = .32$ .

ables rather than to the experimental manipulation of expectancy. One such variable or set of variables is ability or aptitude. Analyses of variance on the pretest and on the three SET subtests gave evidence that there were no significant differences in any ability measure across expectancy conditions. Thus, it is very unlikely that any differences in performance between groups were due to ability differences.

The results supported Hypothesis 1a. As predicted, there was a significant difference of 80.03 to 18.58 between high and low expectancy conditions in perceived effort-pay probabilities ( $F = 448.60$ ,  $df = 1,141$ ,  $p < .0001$ ,  $\omega^2 = .57$ ). There was also a significant interaction between order and expectancy. ( $F = 10.50$ ,  $df = 1,141$ ,  $p < .01$ ). In other words, the magnitude of the difference in effort-pay probability between high and low conditions did depend on the order in which the treatments were received. Nevertheless, as indicated by the comparisons

in Table 3, the differences are significant in the predicted direction regardless of the order of treatment reception.

Hypothesis 1b received no support. Effort in this case was equal to the subject's perception of the extent to which the job input of "trying hard" or "putting out a lot of effort" was present relative to other inputs. There was no significant difference in the perceived amount of effort exerted under the two expectancy conditions ( $F = .54$ ,  $df = 1,135$ ,  $ns$   $\omega^2 = .001$ ). The mean perceived level of effort in the high expectancy condition was .751 whereas in the low condition it was .801. The data in Table 4 may provide some clue as to what actually happened. In the group of subjects which shifted from high to low expectancy (HL), there was no significant difference under the two conditions, as indicated by the  $F$  value in Table 4, which shows the averages across each of the three-day periods for both groups. Furthermore the mean perceived effort in the Low  $\rightarrow$  High group was higher under the low pay-performance contingency condition than under the high. This is diametrically opposite of the predicted direction of differences.

The results obtained with total daily raw performance as the dependent variable provided much clearer support of hypothesis 1c than perceived effort did in the case of 1b. The significant difference between the mean of 95.07 under high expectancy and 76.93 under low expectancy accounted for roughly 33% of the total variance in performance ( $F = 204.62$ ,  $df = 1,141$ ,  $p < .001$ ). In fact, the significant difference in performance under

TABLE 3

MEAN EFFORT-PAY PROBABILITY IN HIGH AND LOW EXPECTANCY CONDITIONS FOR GROUPS SHIFTING FROM HIGH TO LOW AND FROM LOW TO HIGH EXPECTANCY CONDITIONS

Order	Treatment condition	
	High expectancy (HE)	Low expectancy (LE)
High—Low (Order 1)	90.17	18.30
Low—High (Order 2)	71.76	18.80

Note. Individual comparisons are as follows: HE Order 1 vs. LE Order 1,  $F = 208.70$  ( $p < .01$ ); HE Order 2 vs. LE Order 2,  $F = 113.32$  ( $p < .01$ ).

TABLE 4

MEAN PERCEIVED PRESENCE OF EFFORT IN HIGH AND LOW EXPECTANCY CONDITIONS FOR GROUPS SHIFTING FROM HIGH TO LOW AND FROM LOW TO HIGH EXPECTANCY CONDITIONS

Order	Treatment condition	
	High expectancy (HE)	Low expectancy (LE)
High—Low (Order 1)	.84	.76
Low—High (Order 2)	.67	.84

Note. Individual comparisons are as follows: HE Order 1 vs. LE Order 1,  $F = 1.31$  ( $ns$ ); HE Order 2 vs. LE Order 2,  $F = 5.89$  ( $p < .05$ ).

high and low expectancy occurred regardless of the order in which treatment was received, as evidenced by the means in Table 5. At the same time, if task performance had reached asymptote before experimental effects were measured, the difference might have been stronger. The failure to reach asymptote would also account for the significant Order  $\times$  Expectancy interaction ( $F = 13.20$ ,  $df = 1,141$ ,  $p < .01$ ).

The final hypothesis in this series, Hypothesis 1d, deals more explicitly than earlier hypotheses with the changes in the magnitude of variables due to the shift in expectancy condition. In fact, Tables 6 through 8 contain the daily means of each of the three dependent variables already discussed, both in the group which shifted from high to low expectancy conditions and in the group which shifted from low to high. To evaluate the first predicted change, the level of each dependent variable on Day 3 (preshift) must be compared with the level on Day 4 (postshift); to evaluate the changes which occur over time after the shift, the level or mean on Day 4 must be compared with the level on Day 6. The  $ns$  on which the various  $t$  tests in Tables 6 through 8 were based are shown in Table 6.

The first variable in the causal sequence, effort-reward probability, showed most of the changes which were predicted in Hypothesis 1d. As shown in Table 6, there was a significant decrease from Day 3 to Day 4 and from Day 4 to Day 6 in the high to low expectancy group and a significant increase from Day 3 to

TABLE 5

MEAN RAW PERFORMANCE IN HIGH AND LOW EXPECTANCY CONDITIONS FOR GROUPS SHIFTING FROM HIGH TO LOW AND FROM LOW TO HIGH EXPECTANCY CONDITIONS

Order	Treatment condition	
	High expectancy (HE)	Low expectancy (LE)
High—Low (Order 1)	88.90	75.86
Low—High (Order 2)	100.10	77.79

Note. Individual comparisons are as follows: HE Order 1 vs. LE Order 1,  $F = 35.95$  ( $p < .01$ ); HE Order 2 vs. LE Order 2,  $F = 105.23$  ( $p < .01$ ).

TABLE 6

MEANS AND STANDARD DEVIATIONS ACROSS SIX DAYS FOR EFFORT-PAY PROBABILITY

Day	Group					
	High expectancy → low (HL)			Low expectancy → high (LH)		
	$\bar{X}$	$SD$	$n$	$\bar{X}$	$SD$	$n$
1	80.48	25.74	94	19.77	30.58	104
2	88.97	20.41	91	16.58	29.04	102
3	89.57	20.15	89	27.89	38.04	104
4	43.95	22.26	81	70.56	35.03	97
5	21.93	33.77	81	74.51	34.41	82
6	16.35	31.02	74	73.14	23.81	80

Note.  $t$  test comparisons are as follows: HL(Day 3) vs. HL(Day 4),  $t = 14.32$  ( $p < .01$ ); HL(Day 3) vs. HL(Day 6),  $t = 22.09$  ( $p < .01$ ); HL(Day 4) vs. HL(Day 6),  $t = 7.69$  ( $p < .01$ ); LH(Day 3) vs. LH(Day 4),  $t = 19.32$  ( $p < .01$ ); LH(Day 3) vs. LH(Day 6),  $t = 19.15$  ( $p < .01$ ); and LH(Day 4) vs. LH(Day 6),  $t = 1.63$  ( $ns$ ).

Day 4 in the low to high group. The Day 4 to Day 6 difference was not significant.

The case for perceived effort, which is the next variable in the causal sequence, is somewhat weaker than for the other two variables. The results in Table 7 simply confirm the results presented earlier in perceived effort; the only predicted changes over time due to the shift occurred in the high to low group. On the

TABLE 7  
MEANS AND STANDARD DEVIATIONS ACROSS SIX DAYS FOR PERCEIVED PRESENCE OF EFFORT

Day	Group			
	High expectancy → low (HL)		Low expectancy → high (LH)	
	$\bar{X}$	$SD$	$\bar{X}$	$SD$
1	.82	.63	.86	.60
2	.84	.67	.71	.79
3	.92	.71	.97	.63
4	.69	.75	.60	.68
5	.96	.76	.97	.64
6	.71	.88	.52	.76

Note.  $t$  test comparisons are as follows: HL(Day 3) vs. HL(Day 4),  $t = 1.87$  ( $p < .01$ ); HL(Day 3) vs. HL(Day 6),  $t = 4.79$  ( $p < .01$ ); HL(Day 4) vs. HL(Day 6),  $t = .31$  ( $ns$ ); LH(Day 3) vs. LH(Day 4),  $t = 3.60$  ( $p < .01$ ); LH(Day 3) vs. LH(Day 6),  $t = 11.19$  ( $p < .01$ ); and LH(Day 4) vs. LH(Day 6),  $t = 2.40$  ( $p < .05$ ).

TABLE 8  
MEANS AND STANDARD DEVIATIONS ACROSS SIX DAYS  
FOR MEAN HOURLY PERFORMANCE

Day	Group			
	High expectancy → low (HL)		Low expectancy → high (LH)	
	$\bar{X}$	SD	$\bar{X}$	SD
1	18.61	3.58	16.26	3.68
2	21.60	4.25	19.20	4.59
3	23.16	4.62	20.34	5.02
4	19.60	3.71	23.56	5.59
5	18.93	4.45	25.28	4.55
6	17.69	3.32	25.50	5.20

Note.  $t$  test comparisons are as follows: HL(Day 3) vs. HL(Day 4),  $t = 18.53$  ( $p < .01$ ); HL(Day 3) vs. HL(Day 6),  $t = 6.74$  ( $p < .01$ ); HL(Day 4) vs. HL(Day 6),  $t = 2.29$  ( $p < .05$ ); LH(Day 3) vs. LH(Day 4),  $t = -13.68$  ( $p < .01$ ); LH(Day 3) vs. LH(Day 6),  $t = -20.05$  ( $p < .01$ ); and LH(Day 4) vs. LH(Day 6),  $t = -3.58$  ( $p < .01$ ).

other hand, performance dropped significantly in the high to low group and increased significantly in the low to high group from Day 3 to 4 and from Days 4 to 6. These data are shown in Table 8. Thus, there was considerable support for the predictions made in hypothesis 1d.

Table 9 presents the results that were used to test Hypothesis 2a and 2b. These hypotheses amounted to the simple prediction of positive correlations between the two major components of Expectancy  $\times$  Value theory and some de-

pendent variable(s). To test 2a, a multiple regression was carried out with the 13 effort-reward probabilities associated with the 13 job outcomes for a given day as the independent variables and both perceived effort and raw performance for that same day as dependent variables. Hypothesis 2b was tested in the same manner by using the valences for the 13 outcomes as independent variables.

The results of regressing effort-reward probabilities and valence against effort and performance are contained in Table 9. When performance was used as the dependent variable, there was consistent support for both Hypothesis 2a and 2b. In contrast, the correlations with effort produced virtually no support. On only 1 of the 6 days did the multiple correlations between either set of independent variables and effort achieve significance at the .05 level.

## DISCUSSION

It is discouraging that the experimental tests of hypotheses about effort failed to be confirmed. Either the predictions were incorrect or the measurement of effort was inappropriate. One possibility is that self-ratings of the effort variable are invalid. Our use of self-ratings of effort was based, in part, on the successful use of self-ratings in test of the Porter and Lawler model (Porter & Lawler, 1968; Schuster, Clark, & Rogers, 1971). An important difference be-

TABLE 9  
MULTIPLE CORRELATIONS BETWEEN EFFORT-Reward PROBABILITIES, VALENCES,  
AND BOTH PERCEIVED PRESENCE OF EFFORT AND PERFORMANCE

Independent variable <sup>a</sup>	Dependent variable	Day ( $X_i$ )					
		1	2	3	4	5	6
$E_{ij}$ $i = 1, 13$ $j = 1, 6$	Effort Performance	.30 .35*	.24 .43**	.17 .42**	.28 .47**	.41** .49**	.28 .53**
$V_{ij}$ $i = 1, 13$ $j = 1, 6$	Effort Performance	.36* .34*	.25 .34*	.19 .37*	.27 .35*	.27 .37*	.32 .40**

Note.  $n = 187$

\* Independent variable notation is as follows:

$E_{ij} = \text{Effort} - \text{Reward probabilities for the } i = 1, 13 \text{ rewards on } j = 1, 6 \text{ days}$

$V_{ij} = \text{Valence or perceived importance of the } i = 1, 13 \text{ rewards on } j = 1, 6 \text{ days}$

\*\*  $p < .05$

\*\*\*  $p < .01$

tween our measure and the others, however, is that ours involved a multiple paired comparison ranking of nine perceived input variables; thus, each subject's estimate of his own effort is perhaps purely or at least partially ipsative. When such a measurement technique is used, a subject's effort level could differ at two different times on an absolute scale but still show no difference on an ipsative measure if he ranked effort in the same way relative to other input variables. If this occurred for any large proportion of subjects, it could explain our failure to obtain higher mean ratings of perceived effort under the high performance-reward contingency conditions, as well as the absence of many predicted correlations between effort and other variables. Problems with such a measure have been theoretically and empirically demonstrated elsewhere (Clemans, 1966; Hicks, 1970; Knapp, 1964).

Our manipulation of the contingency between pay and performance *did* produce differences in subject's subjective estimates of the relationship between effort and pay; the predicted changes in these estimates over time were also largely supported. This finding supports the hypothesized feedback loop between the objective contingency and the subjective perception of this contingency. The only exception was that subjects in the high → low group did not show significant increases in their estimates of these relationships over the first 3 days. Instead, their estimates reached a maximum immediately and remained there. However, variances of their estimates decreased over the three days, showing that consensus among these subjects may have increased during this period.

Predictions about performance were generally confirmed by the data, with the exception of the failure of subjects in the low → high group to show performance decrements over the first 3 days as expected. The performance increase for this group during the first 3 days suggests that task learning was still occurring, motivated, perhaps, by the subject's perceiving job outcomes (rewards) to be present, which we failed to control adequately. Such outcomes could have maintained or increased the motivation of these subjects to perform on the task even though they correctly perceived that their pay was not related to their performance.

Contrary to prediction, the relative importance attached to money as an outcome increased over time for our subjects; the trend was more pronounced under the high expectancy condition than under low expectancy conditions. On the other hand, the absence of a significant difference in valence of money between expectancy conditions was important because it was necessary for concluding that effort and performance differences were due to expectancy differences and not valence differences.

This investigation did provide evidence that at least one objective performance-reward contingency exerts a causal influence both on an individual's perceptions of this contingency and on his behavior. This evidence adds to the growing body of supporting experimental evidence of this causality obtained almost exclusively in the laboratory. While tests for causality are extremely difficult to carry out in field settings, this study did bridge the gap between laboratory and field by conducting the experiment in a simulated field setting.

Midway through our experiment, we tested inferences about causality by shifting subjects working under each of two performance-reward contingency levels to the other level. Since the change was accompanied by predicted changes in both perceptions and performance, the evidence does support the likelihood of a causal link between the treatment variable and these other variables. This suggests that it may be important to identify environmental conditions first and to examine the way other attitudinal and behavioral variables may intervene between the environmental conditions and observable behavior.

Another strong feature of this study was that it was carried out over a fairly long time period. Most experimental studies have been limited to a few hours. Subjects have only rarely been required to work on more than 1 day. In contrast, our subjects worked for 6 days, 4 hours per day. This longer time period allowed more severe tests of the Expectancy × Value model by affording time for short-term behavioral or attitudinal effects to dissipate. It also provided the opportunity to gather a wealth of information about changes occurring over time.

Even though correlational techniques say

little about causality, they do supplement information about group trends. Unfortunately, measurement problems in the study made interpretation of the correlational data tenuous at best; the method thus added less to our understanding of the data than we had originally hoped.

What is sorely needed is to determine the effects of manipulating objective performance-reward contingencies on (a) the perceived importance of other possible rewards (such as recognition, work itself, etc.) in addition to pay, (b) subjects' perceptions of contingencies between what he does and the likelihood of his receiving these several rewards, (c) subjects' level of satisfaction with rewards, and (d) subjects' estimates of other possible job inputs (e.g., education, previous experience, etc.) in addition to effort. Studies similar to this one should also examine the causal effects that other performance-reward contingencies (such as effort and recognition, performance and relations with co-workers, etc.) may have on both the perceptions of these contingencies and behavior.

#### REFERENCES

- ATKINSON, J. W. Towards experimental analysis of human motivation in terms of motives, expectancies, and incentives. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society*. Princeton: Van Nostrand, 1958.
- BENNET, G. K., & GELINK, M. *The short employment tests*. New York: The Psychological Corporation, 1956.
- CAMPBELL, J. P., DUNNETTE, M. D., LAWLER, E. E., & WEICK, K. E. *Managerial effectiveness: Current knowledge and research needs*. New York: McGraw-Hill, 1970.
- CLEMANS, W. V. An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs*, 1966, 14.
- GRAEN, G. B. Instrumentality theory of work motivation: Some experimental results and suggested modifications. *Journal of Applied Psychology*, 1969, 53 (2, Pt. 2).
- GULLIKSEN, H., & TUCKER, L. R. A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika*, 1961, 26, 173-183.
- HICKS, L. E. Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 1970, 74, 167-184.
- KNAPP, R. R. An empirical investigation of the concurrent and observational validity of an ipsative versus a normative measure of six interpersonal values. *Educational and Psychological Measurement*, 1964, 24, 65-73.
- LEWIN, K. *The conceptual representation and measurement of psychological forces*. Durham, N. C.: Duke University Press, 1938.
- PORTER, L. W., & LAWLER, E. E., III. *Managerial attitudes and performance*. Homewood, Ill.: Irwin-Dorsey, 1968.
- PRITCHARD, R. D., DUNNETTE, M. D., & JORGENSEN, D. O. The effects of perceptions of equity and inequity on worker performance and satisfaction. *Journal of Applied Psychology*, 1972, 56, 75-94.
- SCHUSTER, J. R., CLARK, B., & ROGERS, M. Testing portions of the Porter and Lawler model regarding the motivational role of pay. *Journal of Applied Psychology*, 1971, 55, 187-195.
- STEPHENSON, W. *The Study of behavior: Q-technique and its methodology*. Chicago: University of Chicago Press, 1953.
- TOLMAN, E. C. Purposive behavior in animals and men. New York: Century, 1932.
- VROOM, V. H. *Work and motivation*. New York: Wiley, 1964.
- WEICK, K. E. Organizations in the laboratory. In V. H. Vroom (Ed.), *Methods of organizational research*. Pittsburgh: University of Pittsburgh Press, 1967.

(Received December 6, 1971)

## PREDICTING THE EMERGENCE OF LEADERS USING FIEDLER'S CONTINGENCY MODEL OF LEADERSHIP EFFECTIVENESS<sup>1</sup>

ROBERT W. RICE AND MARTIN M. CHEMERS<sup>2</sup>

*University of Utah*

Eighteen four-man groups participated in a laboratory experiment testing Fiedler's contingency model of leadership effectiveness. Predictions were generated from the model regarding (a) the leadership style (high or low on the least preferred co-worker [LPC] scale) of emergent leaders and (b) the leadership effectiveness of emergent leaders. The attempt to predict leadership style of emergent leaders was unsuccessful. The predictions of leadership effectiveness were accurate and provided support for the model. Sociometric data indicated that low LPC subjects were perceived as more popular and valuable group members than were high LPC subjects.

Fiedler's (1967) contingency model of leadership effectiveness has been one of the most influential theories in the field of leadership research. The model proposes that leadership effectiveness, as reflected by group productivity, is contingent upon the interaction of the leader's orientation and the favorableness of the group task situation. Leadership orientation is measured by Fiedler's "esteem for the least preferred co-worker" (LPC) scale. Fiedler (1967) originally proposed that low LPC leaders are task oriented and primarily motivated toward task achievement while high LPC leaders are relationship oriented and primarily motivated toward establishing rewarding interpersonal relationships. Fiedler (1970) recently modified this interpretation to include secondary motivational systems. Situational favorableness reflects the degree to which the leader has influence and control over the group's activities. Favorableness is determined by the nature of the leader-member relations, the degree of structure inherent in the task, and the leader's position power. The contingency model describes the relationship between leadership orientation (or style) and leadership effectiveness in terms of correlations between

leader LPC and group productivity at different points along the favorableness dimension.

Fiedler (1971b) has reviewed attempts to test the model in both applied and experimental settings. This review maintained that the model has generally been supported by recent research, especially field studies. Despite the supportive findings of these studies, the model is currently the subject of a heated controversy between Fiedler and Graen and his associates (Graen, Alvares, Orris, & Martella, 1970; Graen, Orris, & Alvares, 1971a, 1971b). Graen et al. have charged that the model lacks predictive validity. They maintain that the contingency model represents a post hoc arrangement of research results and that research since the formal exposition of the model (Fiedler, 1964, 1967) has not been generally supportive. Fiedler (1971a) has argued that the data used to substantiate the Graen et al. charges were collected in methodologically faulty experiments. Chemers and Skrzypek (1972) offered exceptionally strong support for Fiedler's defense of the predictive powers of the contingency model. Chemers and Skrzypek conducted an experimental test of all eight octants of favorableness as specified by the model. The obtained leadership effectiveness correlations in this study were strikingly similar to the point predictions generated from the model.

The present study was an attempt to extend the predictive powers of the contingency model. Several field and laboratory studies conducted by Fiedler and his associates have

<sup>1</sup> This research was supported by a grant from the University of Utah Research Committee, "An Interaction Theory Approach to Emergent Leadership," to Martin M. Chemers, principal investigator.

<sup>2</sup> Requests for reprints should be sent to Martin M. Chemers, Department of Psychology, University of Utah, Salt Lake City, Utah 84112.

dealt with emergent leadership (Fiedler, 1954; Fiedler, Meuwese, & Oonk, 1961; Fiedler, O'Brien, & Ilgen, 1969; O'Brien, Fiedler, & Hewett, 1971). However, all of the above studies have been limited to the investigation of effectiveness of emergent leaders. No previous studies of emergent leadership involving the LPC variable have tried to predict which members will emerge as group leaders. The primary purpose of the present study was to determine if the contingency model can be used to predict the leadership style (high LPC or low LPC) of emergent leaders of groups varying in situational favorableness. The present authors proposed that subjects are most likely to emerge as leaders in situations where they can also most effectively fulfill the role of group leader. Hemphill (1961) listed several situational variables that are useful predictors of attempted leadership in emergent leadership situations. Basic to Hemphill's position is the assumption that individuals can and do assess situations in which they are likely to be successful as leaders. High LPC leaders, then, should emerge most often in situations in which high LPC leaders have been shown to be most effective, that is, situations of intermediate favorableness. Low LPC leaders should emerge most often in either very favorable or very unfavorable situations. Implicit in this proposal was the assumption that subjects can somehow recognize situations in which they could lead most effectively.

In addition to testing the ability of the model to predict which subjects would emerge as leaders, the present study also provided an opportunity to further evaluate the ability of the model to predict leadership effectiveness. In order to provide a partial test of the model, the relationship between group productivity and the LPC scores of emergent leaders was examined for two cells of the eight-celled contingency model. Such a test appeared particularly appropriate and meaningful in light of the recent attacks on the model leveled by Graen and his associates.

In order to test the predictive powers of the contingency model, leadership emergence and leadership effectiveness were examined in Octants VI and VIII of the favorableness dimension. Consistent with Fiedler's (1967) de-

scription of the favorableness dimension, Octant VI was characterized by moderately poor leader-member relations, a structured task, and weak leader position power. Octant VIII was characterized by moderately poor leader-member relations, an unstructured task, and weak leader position power. These two octants were particularly appropriate for the present tests of the model because the model predicts that high LPC leaders are most effective in Octant VI and low LPC leaders are most effective in Octant VIII.

In view of the above discussion, the following hypotheses are offered:

1. Hypothesis *a*: Subjects are more likely to emerge as leaders in situations where their leadership styles have been shown to be most effective. More specifically, the emergent leaders in Octant VI are more likely to be high LPC subjects, while the emergent leaders in Octant VIII are more likely to be low LPC subjects.

2. Hypothesis *b*: Rank-order correlations between LPC scores of emergent leaders and group productivity will conform to predictions of the model. Based on Fiedler's (1964, 1967) formal exposition of the model, the predicted correlation for Octant VIII is  $- .43$ . Up to 1967, there had been no studies of Octant VI, but based on interpolation of the curve a correlation of approximately  $.20$  is predicted for Octant VI (Fiedler, 1967, p. 146). Thus, the predicted difference between the leadership effectiveness correlations in Octants VI and VIII is approximately  $.63$ .

## METHOD

### Subjects

The experimental subjects were 72 male undergraduates enrolled in introductory psychology classes at the University of Utah. The subjects volunteered to participate in the experiment and received academic credit in exchange for their participation. The subjects were selected on the basis of their LPC scores. A total of 263 male students completed the LPC scale; subjects scoring in the upper and lower third of the distribution were placed on a list of students eligible to participate in the experiment. All subjects in the experiment were recruited from this list.

The experimental subjects were divided into 18 four-man groups; two high and two low LPC subjects were randomly assigned to each group.

An additional 24 introductory psychology students

(14 male and 10 female) served as raters for the stories written by nine of the experimental groups. These subjects also received class credit in exchange for their participation.

### Design

Nine of the groups operated under the conditions of Octant VI, and the other nine were under the conditions of Octant VIII. Task structure was manipulated while leader-member relations and leader position power were held constant to produce the appropriate levels of favorableness. In Octant VI, leader position power was weak, the task was structured, and leader-member relations were moderately poor. In Octant VIII, leader position power was weak, the task was unstructured, and leader-member relations were moderately poor.

*Position power.* It was assumed that the position power of all emergent leaders would be weak because they had no authority to administer sanctions or rewards to the other group members. They, in fact, possessed no formal leadership status.

*Leader-member relations.* Fiedler (1967, pp. 111-115) proposed that emergent leadership in ad hoc laboratory groups results in "moderately poor" leader-member relations because the group members compete with one another for the leadership position. On the basis of Fiedler's analysis, it was assumed that the relations would be moderately poor for groups working on both the structured and unstructured tasks.

*Task structure.* The nine groups in Octant VI worked on a structured task while the nine groups in Octant VIII worked on an unstructured task.

The structured task required the group to draw the front view of a house. The subjects were given a drawing of the house with dimensions given in metric units. Two conversion tables allowed the subjects to convert the metric units into scaled inches by following a two-step transformation process requiring the use of both conversion tables. The groups were required to draw the house in scaled inches. This task was patterned after the structured task used by Chemers and Skrzypek (1972). The groups were allowed 45 minutes to complete the task. Two groups completed the task in the allotted time. The house-plan task was considered highly structured because the goal was clear, the correctness of each step was easily verified, the number of alternative goal paths was severely limited, and there was only one correct solution. Group productivity for this task was defined as the average number of correct lines drawn per minute.

The unstructured task required the group to write two original stories based on a group discussion of a single Thematic Apperception Test (TAT) picture. This was the same task used by Fiedler et al. (1961). The time limit for this task was also 45 minutes. The TAT task was considered quite unstructured because the goal was vague and ambiguous, the correctness of the solution was difficult to objectively verify, the number of alternative goal paths was virtually unlimited, and there was no single "correct" solution.

Group productivity for this task was based on the summed ratings of 24 raters. The raters assessed each story on a 6-point scale of "overall quality." The raters were instructed to consider writing style and clarity, comprehension, interest, and creativity in their ratings of overall quality. The reliability coefficient for the summed ratings of overall quality was .795 (Guilford, 1965, pp. 297-300). The productivity score for each group was determined by summing the ratings of the two stories written by each group. Although it was not possible to determine the validity of the ratings, it was encouraging to discover that the rank-order correlation between the summed ratings and the number of words in each story was -.08. This suggested that the raters were attending to the content of the stories and not simply to the length of each story.

### Procedure

The experimental sessions were conducted on weekday evenings in large classrooms. Several groups were run simultaneously with sufficient space between groups to prevent eavesdropping.

Instructions stressed that the final product was to reflect the efforts of the entire group. To promote group activity and minimize independent task activity, each group was given the minimal amount of supplies (rulers, pencils, paper, etc.) necessary to complete the task. Further, the subjects were prohibited from using their own supplies. It was hoped that these precautions and instructions would preclude group members from working as individuals on the tasks.

Following the 45-minute experimental session, each subject completed a sociometric questionnaire. Each subject indicated which group member had emerged as the leader. If leadership was shared by two or more group members, the subjects were required to estimate the percentage of total leadership exercised by each nominated leader. The subject with the highest nomination score in each group was considered to be the emergent leader. The subjects were also asked to provide the following information: (a) the group members with whom they most and least enjoyed working, (b) the group members whom they would prefer as leader and as co-worker for a similar task in the future, (c) the most valuable member of the group, and (d) the socioemotional leader of the group.

## RESULTS

### Emergence of Leaders

Hypothesis *a* proposed that more high LPC subjects would emerge as leaders in Octant VI while more low LPC subjects would emerge as leaders in Octant VIII. This hypothesis was not supported. In Octant VI, four high LPC subjects and five low LPC subjects emerged as group leaders. In Octant VIII, six of the emergent leaders had low LPC scores while three had high LPC scores. A  $2 \times 2$  chi-

square test of these frequencies failed to reach an acceptable level of significance ( $\chi^2 < 1$ ,  $df = 1$ ).

### *Leadership Effectiveness*

Hypothesis *b* proposed that the correlations between emergent leader LPC and group productivity would conform to the point predictions based on the contingency model for Octants VI and VIII. The obtained correlations were .30 and -.40 for Octants VI and VIII, respectively. Although neither of these correlations reached conventional levels of significance, they compared favorably with the point predictions of .20 and -.43. The obtained absolute difference between the leadership effectiveness correlations in Octants VI and VIII was .70; this result compared favorably with the predicted difference of .63.

The close correspondence between the predicted and obtained correlations provided some support for Hypothesis *b*.

### *Sociometric Questionnaire Data*

Because most LPC research has dealt exclusively with the behavior and attitudes of leaders, it was not possible to generate hypotheses concerning the relationship between member LPC and the sociometric data. However, the questionnaire yielded an interesting and unexpected pattern of prominence and popularity as a function of member LPC.

The frequency of nomination for each category was first analyzed by means of  $2 \times 2$  chi-square tests (High LPC and Low LPC Nominees  $\times$  Octants VI and VIII). None of the chi-square values were significant because the frequencies did not differ between Octants VI and VIII. The nomination frequencies were then collapsed across the two octants in order to reduce the four-celled tables into two cells (high LPC nominees and low LPC nominees). Single sample chi-square tests (Siegel, 1956, pp. 42-47) were used to test the significance of differences in the frequency of nominating high and low LPC subjects for each category. For each category the expected values were based on the null hypothesis that frequency of nomination does not differ as a function of the LPC score of the nominee. The sample size for these questions often varied because some subjects nominated more than

one member for a single category (e.g., their most enjoyed co-worker). On the other hand, some subjects refused to nominate any group member for certain categories (e.g., their least enjoyed co-worker).<sup>3</sup>

*Future leader.* Significantly more low LPC subjects were nominated as "leader for a similar task sometime in the future" ( $\chi^2 = 13.84$ ,  $df = 1$ ,  $p < .001$ ). Fifty-three of the nominees (25 in Octant VI and 28 in Octant VIII) were low LPC subjects, while only 21 (11 in Octant VI and 10 in Octant VIII) were high LPC subjects. Apparently low LPC subjects were strongly preferred as leaders for future groups.

*Future co-worker.* Significantly more low LPC subjects were nominated as preferred co-worker on a similar task in the future ( $\chi^2 = 12.24$ ,  $df = 1$ ,  $p < .001$ ). There were 61 low LPC subjects (31 in Octant VI and 30 in Octant VIII) and only 28 high LPC subjects (13 in Octant VI and 15 in Octant VIII) nominated as preferred future co-workers. Apparently low LPC subjects were perceived as more desirable co-workers.

*Most valuable member.* Consistent with the results regarding future leaders and future co-workers, low LPC subjects were nominated more frequently as "the single, most valuable member of the group" ( $\chi^2 = 5.88$ ,  $df = 1$ ,  $p < .02$ ). Forty-eight low LPC subjects (22 in Octant VI and 26 in Octant VIII) and 27 high LPC subjects (14 in Octant VI and 13 in Octant VIII) were nominated as the most valuable group member. Low LPC subjects were perceived as more important contributors to group success than were high LPC subjects.

*Most and least enjoyed co-workers.* Significantly more low LPC subjects were nominated

<sup>3</sup> There was some question regarding the treatment of nomination data for subjects who nominated more than one group member for a given category. The analyses reported in the text of the paper included all multiple nominations in the calculation of chi-square values. To determine if inclusion of multiple nominations biased the nomination pattern in any way, all chi-square analyses of the sociometric data were repeated using only the nomination data from subjects who nominated a single group member for each category. For every category, the results of the analyses with the multiple nominations deleted paralleled the results of the analyses in which the multiple nominations were included. No changes in direction or significance level of results were found.

as most enjoyed co-worker ( $\chi^2 = 3.86, df = 1, p < .05$ ). Fifty-one low LPC subjects (27 in Octant VI and 24 in Octant VIII) and 33 high LPC subjects (14 in Octant VI and 19 in Octant VIII) were nominated as the most enjoyed co-worker. Significantly more high LPC subjects were nominated as the least enjoyed co-worker ( $\chi^2 = 6.67, df = 1, p < .01$ ). Forty high LPC subjects (21 in Octant VI and 19 in Octant VIII) and 20 low LPC subjects (9 in Octant VI and 11 in Octant VIII) were nominated as the least enjoyed co-worker. The responses to these two questions indicate that low LPC subjects were perceived as more popular and enjoyable co-workers than were high LPC subjects.

*Socioemotional leader.* The questionnaire also requested the subjects to indicate if there was anyone in the group that fit a role description of a socioemotional leader. There were no significant differences in the frequency of nomination for socioemotional leader as a function of LPC scores ( $\chi^2 < 1, df = 1$ ). There were 28 high LPC subjects (14 in both Octants VI and VIII) and 32 low LPC subjects (20 in Octant VI and 12 in Octant VIII) nominated as socioemotional leaders. Apparently LPC was not systematically related to perceived socioemotional leadership. The same subjects did not usually fulfill the roles of both socioemotional leader and task leader for a given group. Only two of the 18 groups nominated the same subject as both the task and socioemotional leader. The other 16 groups showed a pattern of role differentiation, that is, the roles of the task leader and the socioemotional leader were filled by different group members.

In order to determine if the pattern of nomination was associated with the LPC score of the nominator, a second series of single sample chi-square tests was conducted. The nominations made by high LPC subjects and by low LPC subjects were examined separately. For three categories (future leader, socioemotional leader, and most valuable member) the expected values were calculated on the assumption that each nominator was equally likely to nominate a high LPC subject or a low LPC subject, that is, each nominator could choose from two high LPC members or two low LPC members, including himself. Ex-

amination of the nomination data supported this assumption; subjects frequently nominated themselves for these categories. For three other categories (future co-worker, most enjoyed co-worker, and least enjoyed co-worker) the expected values were calculated on the assumption that a subject could not reasonably nominate himself for these categories. Assuming that the nominator excludes himself, there are two members of opposite LPC and one member of his own LPC for him to nominate, for example, the probability that a low LPC subject nominates a high LPC subject is two to three and the probability that a low LPC subject nominates a low LPC subject is one to three. Examination of the nomination data supported this assumption; subjects never nominated themselves for these categories.

*Future leader.* High LPC subjects nominated 26 low LPC subjects and 11 high LPC subjects ( $\chi^2 = 6.08, df = 1, p < .02$ ). Low LPC subjects nominated 27 low LPC subjects and 10 high LPC subjects ( $\chi^2 = 7.81, df = 1, p < .01$ ). Both high LPC nominators and low LPC nominators nominated significantly more low LPC subjects as future leader.

*Most valuable member.* High LPC subjects nominated 22 low LPC subjects and 13 high LPC subjects ( $\chi^2 = 2.31, df = 1, p < .20$ ). Low LPC subjects nominated 26 low LPC subjects and 14 high LPC subjects ( $\chi^2 = 3.60, df = 1, p < .10$ ). Both high LPC nominators and low LPC nominators nominated more low LPC subjects as most valuable member, but the chi-square values did not reach conventional levels of significance.

*Socioemotional leader.* High LPC subjects nominated 16 low LPC subjects and 12 high LPC subjects ( $\chi^2 < 1, df = 1$ ). Low LPC subjects nominated 16 low LPC subjects and 16 high LPC subjects ( $\chi^2 < 1, df = 1$ ). Neither high LPC nominators nor low LPC nominators showed any preference in their nominations for socioemotional leader.

*Future co-worker.* High LPC subjects nominated 38 low LPC subjects and 4 high LPC subjects ( $\chi^2 = 10.71, df = 1, p < .01$ ). Low LPC subjects nominated 23 low LPC subjects and 24 high LPC subjects ( $\chi^2 = 5.13, df = 1, p < .05$ ). Both high LPC nominators and low LPC nominators nominated significantly more low LPC subjects as future co-workers than

would be expected based on the two-to-three and one-to-three probabilities.

*Most enjoyed co-worker.* High LPC subjects nominated 33 low LPC subjects and 11 high LPC subjects ( $\chi^2 = 1.38$ ,  $df = 1$ ,  $p < .30$ ). Low LPC subjects nominated 18 low LPC subjects and 22 high LPC subjects ( $\chi^2 = 2.30$ ,  $df = 1$ ,  $p < .20$ ). Both high LPC nominators and low LPC nominators nominated more low LPC subjects as most enjoyed co-worker than expected based on the two-to-three and one-to-three probabilities, but the chi-square values did not reach conventional levels of significance.

*Least enjoyed co-worker.* High LPC subjects nominated 13 low LPC subjects and 14 high LPC subjects ( $\chi^2 = 4.17$ ,  $df = 1$ ,  $p < .05$ ). Low LPC subjects nominated 7 low LPC subjects and 26 high LPC subjects ( $\chi^2 = 2.17$ ,  $df = 1$ ,  $p < .20$ ). Both high LPC nominators and low LPC nominators nominated more high LPC subjects as least enjoyed co-worker than expected based on the two-to-three and one-to-three probabilities, but the difference was significant for only the high LPC nominators.

This series of analyses clearly demonstrated that nomination patterns were the same for both high LPC nominators and low LPC nominators.

A two-way analysis of variance with unequal cell frequencies (Winer, 1962, pp. 241-244) was used to test for an interaction effect in the nomination of emergent leaders. The "percentage of leadership" scores for subjects nominated as emergent leaders were analyzed in a  $2 \times 2$  design (High LPC and Low LPC Nominators  $\times$  High LPC and Low LPC Nominees). All main effects and interactions failed to reach conventional levels of significance. The insignificant interaction effect indicated that the LPC scores of nominators did not significantly influence their nominations of emergent leaders.

## DISCUSSION

The attempt to use the contingency model to predict which type of subjects would emerge as group leaders was not successful. Apparently, there is no simple relationship between leadership effectiveness and leader

emergence when working within the framework of Fiedler's model. Based on the present findings, the model appears to be most appropriate for the prediction of leadership effectiveness and cannot be used to predict leadership emergence. There are numerous plausible explanations for the failure to predict the emergence of leaders using the model. However, the simplest and most straightforward explanation is that subjects simply do not know or recognize those situations in which their individual leadership style would be most effective.

Although the model did not successfully predict which subjects would emerge as group leaders, the leadership effectiveness predictions appeared to be accurate. Unfortunately, the present authors know of no statistical procedure for testing the accuracy of a point prediction such as those generated from the contingency model. However, by inspection, the obtained correlations for Octants VI and VIII appear to offer strong support for the ability of the contingency model to predict leadership effectiveness. The close correspondence between the predicted and obtained correlations of the present study, in conjunction with the strong support for the model offered by Chemers and Skrypek (1972), lead the present authors to reject the Graen et al. (1970) contention that the contingency model lacks predictive validity.

The sociometric questionnaire indicated that low LPC subjects were generally more popular and highly valued than were high LPC subjects. The second series of single sample chi-square analyses indicated that this pattern held for both high LPC nominators and low LPC nominators. These results may reflect the subjects' general perception of the experimental situation. The experiment was presented as a task situation, and the subjects probably perceived strong task demands to their situation. The behavior pattern and personality of low LPC subjects may have been viewed more positively from such a task demand perspective. According to Fiedler's (1970) most recent interpretation of the relationship between leader behavior and LPC, leaders should engage in behavior reflecting their primary motivational goals under the

relatively adverse conditions of Octants VI and VIII. If Fiedler's new hypothesis can be applied to the present findings concerning all group members, then the strong task orientation of low LPC subjects may account for the popularity and value attributed to them by their fellow group members. The task orientation of the low LPC subjects may have simply appeared more appropriate to the task demands of the experiment than the relationship orientation of the high LPC subjects. However, such an application of Fiedler's theory must be made with caution since Fiedler has been concerned specifically with leader behavior and the present data are based on all group members.

The finding that nomination for socioemotional leader did not differ as a function of the LPC score of the nominee was enlightening. It might be expected that high LPC subjects would tend to become socioemotional leaders more frequently because of their strong relationship orientation. However, it must be realized that the high LPC subject is not necessarily interested in the socioemotional leader's function of maintaining good social relations within the group as a unit. Rather, the high LPC subject is concerned with achieving and maintaining gratifying and rewarding relationships between himself and other members of the group. With this distinction in mind, it appears reasonable that high LPC subjects need not be more frequently nominated as socioemotional leaders.

The present finding that different subjects filled the roles of socioemotional leader and task leader is consistent with the theories of role differentiation proposed by Bales (1958) and Slater (1955). Bales reported that groups often have individual members who serve as task or socioemotional specialists; it is exceptional that a single "great man" can fulfill both roles. The present findings support the contention that a single individual is seldom able to fulfill both roles.

#### REFERENCES

- BALES, R. F. Task roles and social roles in problem solving groups. In E. E. Maccoby, T. M. Newcomb, & E. L. Hartley (Eds.), *Readings in social psychology*. (3rd ed.) New York: Holt, Rinehart & Winston, 1958.
- CHEMERS, M. M., & SKRZYPEK, G. J. An experimental test of the contingency model of leadership effectiveness. *Journal of Personality and Social Psychology*, 1972, 24, 172-177.
- FIEDLER, F. E. Assumed similarity measures as predictors of team effectiveness. *Journal of Abnormal and Social Psychology*, 1954, 49, 381-388.
- FIEDLER, F. E. A contingency model of leadership effectiveness. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 1. New York: Academic Press, 1964.
- FIEDLER, F. E. *A theory of leadership effectiveness*. New York: McGraw-Hill, 1967.
- FIEDLER, F. E. Personality, motivational systems, and behavior of high and low LPC persons. (Tech. Rep. 70-12) Seattle: University of Washington, Organizational Research Laboratory, September 1970.
- FIEDLER, F. E. Note on the methodology of the Graen, Orris, and Alvares studies testing the contingency model. *Journal of Applied Psychology*, 1971, 55, 202-204. (a)
- FIEDLER, F. E. Validation and extension of the contingency model of leadership effectiveness: A review of empirical findings. *Psychological Bulletin*, 1971, 76, 128-148. (b)
- FIEDLER, F. E., MEUWESE, W., & OONK, S. An exploratory study of group creativity in laboratory tasks. *Acta Psychologica*, 1961, 18, 100-119.
- FIEDLER, F. E., O'BRIEN, G., & ILGEN, D. The effect of leadership style upon the performance and adjustment of volunteer teams operating in a stressful foreign environment. *Human Relations*, 1969, 22, 503-514.
- GRAEN, G., ALVARES, K., ORRIS, J., & MARTELLA, J. Contingency model of leadership effectiveness: Antecedent and evidential results. *Psychological Bulletin*, 1970, 74, 285-296.
- GRAEN, G., ORRIS, J., & ALVARES, K. Contingency model of leadership effectiveness: Some experimental results. *Journal of Applied Psychology*, 1971, 55, 196-201. (a)
- GRAEN, G., ORRIS, J., & ALVARES, K. Contingency model of leadership effectiveness: Some methodological issues. *Journal of Applied Psychology*, 1971, 55, 205-210. (b)
- GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1965.
- HEMPHILL, J. K. Why people attempt to lead. In L. Petrullo & B. M. Bass (Eds.), *Leadership and interpersonal behavior*. New York: Holt, Rinehart & Winston, 1961.
- O'BRIEN, G., FIEDLER, F., & HEWETT, T. The effects of programmed culture training upon the performance of volunteer medical teams in Central America. *Human Relations*, 1971, 24, 209-231.
- SIEGEL, S. *Nonparametric statistics*. New York: McGraw-Hill, 1956.
- SLATER, P. Role differentiation in small groups. *American Sociological Review*, 1955, 20, 300-310.
- WINER, B. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

(Received October 26, 1971)

## SOME INTERACTIONS BETWEEN PERSONALITY VARIABLES AND MANAGEMENT STYLES

KENNETH E. RUNYON<sup>1</sup>

*Northern Arizona University*

The present study investigated the interaction between management style and the personality variable, "locus of control," on workers' "satisfaction with supervision" and "job involvement" among hourly employees of a major, multiplant chemical company. Satisfaction with supervision was found to be a function of the interaction between management style and employee internality. Job involvement was found to be related directly to employee internality, with the interaction of management style and employee internality having a negligible effect on this dependent variable.

The purpose of this study is to investigate the interaction between management style and the personality variable "locus of control" on the attitudes of employees toward their immediate supervisor and toward their work. Previous studies of management style have concentrated on the effects of autocratic versus participatory management on employee attitudes in a variety of industrial settings; however, the interaction between management style and employee personality has been largely neglected.

The personality variable employed in this study is Rotter's concept of "locus of control" (Rotter, 1966). "Locus of control" refers to a generalized belief that a person can or cannot control his own destiny. This belief arises from social learning, and is rooted in general principles of reinforcement. Within the context of social learning, it is argued that individuals receive reinforcement under varying conditions. If an individual perceives a reinforcement as being contingent upon his own actions, this is termed a belief in "internal" control. If an individual perceives a reinforcement as being contingent upon outside forces, it is termed a belief in "external" control. Depending on one's life history, a person builds up generalized expectancies or beliefs concerning the nature of the reinforcements he receives. The generalized expectancies of "internal versus external" control have functional properties that make them an important personality variable.

It is a general hypothesis of this study that individuals who see themselves as Internals and those who see themselves as Externals will react differently to styles of supervision differentiated along a directive-participative continuum.

### DEPENDENT VARIABLES AND SPECIFIC HYPOTHESES

#### *General Considerations*

This study examines the effects of the interaction of management style and worker personality on two dependent variables: (a) satisfaction with supervision and (b) job involvement. These particular variables were selected because it was believed, judgmentally, that they would be more responsive to differences in supervisory style than some of the more complex variables that have been used by other researchers in the field. The conceptual model underlying this rationale is the Likert (1967) model that postulates three sets of variables relating to worker behavior: (a) causal variables (supervisory style); (b) intervening variables (worker attitude); and (c) end result variables (worker behavior as it pertains to productivity, increased sales, etc.). In this model, satisfaction with supervision and job involvement are conceived of as intervening variables.

#### *Satisfaction with Supervision*

Satisfaction with supervision is used as a global measure of the respondent's reaction to the style of management under which he works. No assumption is made that satisfaction with supervision, as such, has profound implications for worker effectiveness or productivity. It is assumed, however, that

<sup>1</sup> Requests for reprints should be sent to Kenneth E. Runyon, College of Business Administration, Northern Arizona University, Box 5736, Flagstaff, Arizona 86001.

dissatisfaction with supervision is a contributing factor in organizational turnover. Vroom (1969) has summarized a number of studies supporting this assumption. Thus, satisfaction with supervision is seen as a factor in the organization-employee relationship that may influence employee behavior in ways which are advantageous or detrimental to organizational welfare.

The first hypothesis of this study involves satisfaction with supervision:

*Hypothesis 1:* The more individuals see themselves as Internals, the greater will be their satisfaction with participative management and vice versa. Hypothesis 1 predicts essentially different reactions to managerial style, depending upon the degree of internality present in the employee. As internality increases, the employee should perceive himself as being better able to control his own destiny. Consequently, he should respond positively to the freedom for personal initiative and responsibility that is characteristic of participative management. In contrast, as internality decreases, the employee should find participative management frustrating and insufficiently structured. In this instance, he should respond by expressing a preference for a more directive management style.

#### *Job Involvement*

Dubin (1956) reports a study of the central life interests of industrial workers in three plants in the midwest. One of his primary findings was that "Work is no longer a central interest for workers. These life interests have moved out into the community [p. 140]." Defining central life interests as "... the expressed preference for a given locale or situation in carrying out an activity [p. 134]," Dubin notes that the traditional assumption that work is of central importance to adults in the Western world may no longer be justified. He further suggests that management efforts to center primary human relationships in work through such devices as participant management and group dynamics have not been remarkably successful.

Taking their lead from Dubin (1956), Lodahl and Kejner (1965) developed the concept of job involvement as the degree to which a person's work performance affects his self esteem, and developed a "job involvement" scale to mea-

sure the degree to which this characteristic is present in employees. On the basis of their findings with this scale, Lodahl and Kejner suggest that job involvement tends to be stable over time and is relatively independent of situational factors.

A contrasting point of view is one that focuses on situational factors as the primary source of job involvement. Lawler and Hall (1970) attribute this point of view to Vroom (1969) who "... has suggested that job factors can influence the degree to which an employee is involved in his job, although he presents little data about the impact of job factors on job involvement [p. 307]."

In an effort to clarify the concept of job involvement, and distinguish it from "job satisfaction" and "intrinsic motivation," Lawler and Hall (1970) applied factor analysis to measures of these variables. On the basis of study of 291 scientists in 22 research and development laboratories, they concluded that job involvement, job satisfaction, and intrinsic motivation were factorially independent and relatively distinct variables. They also suggested:

Perhaps the most realistic view of job involvement is that it is a function of an individual-job characteristic interaction. People probably do differ as a function of their backgrounds and personal situations in the degree to which they are likely to become involved in their job. However, it is also probably true that other things being equal more people will become involved in a job that allows them control and a chance to use their abilities than will become involved in jobs that are lacking in these characteristics [p. 311].

Against this background, a general hypothesis of this study is that the extent of job involvement will vary with the degree of internality present in employees.

*Hypothesis 2:* The more closely supervision approaches participative management, the greater will be the job involvement of individuals who see themselves as Internals.

By Hypothesis 2, the work situation is seen as a way of demonstrating competence for the Internal and, as the work environment permits such demonstration, the degree of job involvement will increase. Under conditions of directive management, where opportunities to demonstrate personal competence are decreased, the level of job involvement should decline.

*Hypothesis 3:* Regardless of style of supervision, job involvement for the individual who perceives himself as an External will be low. This hypothesis suggests that the idea of the work situation as a way of demonstrating competence is a nonsequitur for the External. In a world that is controlled by fate, or luck or by powerful others, the opportunity to demonstrate personal competence in the work situation is virtually nonexistent. Without the basic belief that one can perform well, the work situation, under any supervisory structure, offers little possibility of generating self esteem. Consequently, work involvement should not be a major consideration in the External's psychological life.

Hypotheses 2 and 3 recognize the tentative conclusion of Lawler and Hall (1970) that people probably do differ as a function of their background in the degree to which they are likely to become involved in their jobs. It also recognizes their suggestion that situational factors may also play a role in job involvement. Taken together, Hypotheses 2 and 3 suggest that, under appropriate supervision, Internals will become job involved, whereas Externals are not apt to become job involved under any conditions.

## METHOD

### Subjects

Subjects participating in the study consisted of 110 hourly employees in the manufacturing, packaging, yard and maintenance departments of an urban plant of a large, multilocation chemical company. They were members of 18 supervisory groups. Both subjects and groups were chosen by means of a table of random numbers. Subjects ranged in age from 21 to 64, with a median age of 50; length of service ranged from 0.2 to 47.0 with a median of 22 years. Length of time in their supervisory groups ranged from 0.1 to 15.0 years, with a median of 3.5 years.

### Test Instruments

Data were gathered by means of paper-and-pencil questionnaires. Four separate questionnaires were used; the questionnaires covered the following areas:

*Style of management.* This measure consisted of seven Likert type scales developed specifically for this study. The scales were designed to cover the following areas of supervisory behavior: (a) supervisory consultation with subordinates concerning decisions involving the subordinate's job; (b) willingness on the part of the supervisor to listen to, and seek the opinions of subordinates on matters concerning their work; and (c) supervisory encouragement to show initiative and assume responsibility. The scales were subjected to a generalizability study (Gleser, Cronbach, & Rajarat-

nam, 1965) on a pilot basis in order to see how well one could generalize from specific observations of supervisory behavior to an hypothesized universe of such observations. The "wanted" variation due to differences in supervisory behavior far exceeded the "unwanted" variance in the total scale attributable to idiosyncratic responses of subjects to specific scales. The obtained generalizability coefficient was .90.

*Locus of control.* Twenty-six of the 29 items on the I-E scale were used to measure the internal-external dimension of personality. Three items pertaining to school behavior were dropped because, judgmentally, they were deemed inappropriate in view of the subjects' ages and backgrounds. Biserial item correlation with total score (with that item removed) are moderate but consistent, with most items falling in the .20 to .30 range. Split-half reliability and test-retest reliability are consistent and moderately high, with  $r_s$  in the .65 to .70 range. A summary of studies on scale reliability and its construct validity has been reported by Rotter (1966).

*Work involvement.* The short form of the Lodahl and Kejner job involvement scale was used for this measure. The split-half reliability of the short form is estimated at .73 and the correlation of the short form with the original scale is .87. A review of studies on reliability as well as on discriminant and correlational validity has been reported by Lodahl and Kejner (1965).

*Satisfaction with supervision.* Satisfaction with supervision was measured by a single question in which subjects were asked to indicate their degree of satisfaction on a 7-point scale, with possible responses ranging from 1 (most negative) through 4 (neutral) to 7 (most positive).

*Procedure.* All scales were administered to subjects in groups of 12 to 18 persons. They were asked to identify themselves by a number that had been assigned to each supervisory group. In each supervisory group thus identified, three alternate members were given the style of management scale. The remaining scales were given to the remaining members of each supervisory group. This procedure was used to provide assurance that personality characteristics of subjects who filled out the I-E scale would not influence (i.e., confound) the management style ratings of the supervisors for whom they worked. The individual questionnaires were identified only in terms of supervisory group, so that, throughout, the anonymity of subjects was protected.

## RESULTS

### Distribution of the Independent Variables

The ratings used to describe the style of management for each supervisory group were obtained by taking the arithmetic mean of the three ratings made of each supervisor. Possible ratings on this basis ranged from 7 (directive supervision) to 35 (participative supervision). Nine supervisors received ratings that ranged from 12 to 20; the other nine supervisors received ratings that ranged from 26 to 32.

The I-E scale was filled out by 54 employees—three members in each of 18 supervisory groups. The possible range of scores was from 0 (Internal) to 20 (External). The obtained distribution ranged from 2 to 16, with an arithmetic mean of 8.37 and a standard deviation of 3.68. While these parameters do not appear to be inconsistent with those of other populations sampled (Rotter, 1966), it should be noted that three items were omitted from the I-E scale in this study because they were deemed inappropriate.

#### *Relationship of Independent Variables*

In order to test for the independence of the two major variables, a scattergram of I-E scores and management style ratings was made. Visual inspection of this scattergram showed no apparent systematic relationship between the two variables. This observation was confirmed by the computation of the chi-square test for two independent samples. The obtained chi-square value of .11 ( $df = 2$ ) indicated that the hypothesis of independence was accepted with a  $p$  value between .90 and .95.

*Satisfaction with supervision.* Hypothesis 1 relates satisfaction with supervision to management style and locus of control. Essentially, it states that Internals will be more satisfied under participative management, whereas Externals will be more satisfied under directive supervision.

Table 1 gives the  $F$  values and the levels of significance for the main effects of management style and locus of control on satisfaction with supervision. From this table, it can be seen that management style alone, as well as the interaction of the two independent variables exert an effect on satisfaction with supervision that is statistically significant beyond the .01 level.

TABLE 1  
ANALYSIS OF VARIANCE FOR MAIN EFFECTS OF INDEPENDENT VARIABLES ON SATISFACTION WITH SUPERVISION

Source	df	MS	F
Factor A (management style)	1	9.99	7.57*
Factor B (locus of control)	2	.50	.38
AB (interaction)	2	15.09	11.43**
Error (within cell)	48	1.32	
Total	53	1.97	

\*  $p < .01$ .

\*\*  $p < .001$ .

The nature of these effects can be seen in Table 2, which shows the mean satisfaction scores for each of the test conditions. From this table it can be seen that the mean satisfaction score for Internals under participative management (5.44) is higher than the mean satisfaction score for Internals under directive management (2.89). Conversely, the mean satisfaction score of Externals under directive management (4.75) is higher than the mean satisfaction score for Externals under participative management (3.67).

When the interaverage comparisons in Table 2 are subjected to the Neuman-Keuls test (Winer, 1962, p. 309), the following statements can be made about the findings:

- Under participative management, the satisfaction of Internals is significantly greater than that of Externals ( $\bar{AB}_{11}$  vs.  $\bar{AB}_{13}$ ,  $p < .01$ ).
- Under conditions of directive management, the satisfaction of Externals is significantly greater than that of Internals ( $\bar{AB}_{21}$  vs.  $\bar{AB}_{23}$ ,  $p < .01$ ).
- Internals, under participative management, exhibit greater satisfaction with super-

TABLE 2  
MEAN SATISFACTION SCORES FOR EACH OF THE TEST CONDITIONS

Style of management (A)	Locus of control (B)		
	(b <sub>1</sub> ) Internal (2-6)	(b <sub>2</sub> ) Intermediate (7-10)	(b <sub>3</sub> ) External (11-16)
(a <sub>1</sub> ) Participative (26-32)	$\bar{AB}_{11} = 5.44$	$\bar{AB}_{12} = 5.00$	$\bar{AB}_{13} = 3.67$
(a <sub>2</sub> ) Directive (12-20)	$\bar{AB}_{21} = 2.89$	$\bar{AB}_{22} = 3.90$	$\bar{AB}_{23} = 4.75$

TABLE 3

ANALYSIS OF VARIANCE FOR MAIN EFFECTS OF INDEPENDENT VARIABLES ON WORK INVOLVEMENT SCORES

Source	df	MS	F
Factor A (management style)	1	49.09	3.23*
Factor B (locus of control)	2	281.97	18.57**
AB (interaction)	2	.59	.04
Error	48	15.18	
Total	53	25.34	

\*  $p = .08$ .

\*\*  $p < .001$ .

vision than Internals under directive supervision ( $\bar{AB}_{11}$  vs.  $\bar{AB}_{21}$ ,  $p < .01$ ).

4. Externals, under directive management, exhibit greater satisfaction with supervision than Externals under participatory management ( $\bar{AB}_{13}$  vs.  $\bar{AB}_{23}$ ,  $p = .05$ ).

These findings are in direct support of Hypothesis 1.

*Work involvement.* Hypotheses 2 and 3 relate work involvement to management style and locus of control. Hypothesis 2 states that the work involvement of Internals will be directly related to the amount of participation afforded by the management style under which they work. In contrast to Internals, Hypothesis 3 states that Externals will evidence a low degree of job involvement, regardless of the management style to which they are subjected.

Table 3 shows the *F* values and levels of significance for the main effects of managerial style and locus of control on work involvement. From this table, it can be seen that the personality variable, locus of control, has a major effect on work involvement,  $p < .01$ . Management style may also have some influence on work involvement ( $p = .08$ ), but

the interaction effect of the two major variables on work involvement is negligible.

The specific nature and direction of these effects can be seen in Table 4, which shows the mean work involvement scores for each of the test conditions. From this table it can be seen that the mean work involvement scores increase as one moves from External to Internal, under both management styles. Similarly, although to a lesser extent, mean work involvement scores increase as one moves from directive to participative management in each of the personality categories.

When the interaverage comparisons in Table 4 are subjected to the Neuman-Keuls test, the following statements can be made about the findings:

1. Internals exhibit significantly more job involvement than Externals under both participatory and directive supervision ( $\bar{AB}_{11}$  vs.  $\bar{AB}_{13}$ ,  $p < .01$ ;  $\bar{AB}_{21}$  vs.  $\bar{AB}_{23}$ ,  $p < .01$ ).

2. Job involvement tends to be greater under participatory management than under directive management, but the differences are not statistically significant ( $\bar{AB}_{11}$  vs.  $\bar{AB}_{21}$ ;  $\bar{AB}_{12}$  vs.  $\bar{AB}_{22}$ ;  $\bar{AB}_{13}$  vs.  $\bar{AB}_{23}$ ).

The findings on job involvement only partially support the hypotheses of the study. Hypothesis 3, which states that the work involvement of Externals would tend to be low, regardless of management style, is supported by the data. Hypothesis 2 is not well supported; while the job involvement of Internals is slightly greater under participative supervision than under directive supervision, job involvement is relatively high in both cases. A relatively high correlation ( $r = -.64$ ) between job involvement and the I-E measure was unanticipated in the hypotheses. Thus, the findings suggest that work

TABLE 4  
MEAN WORK INVOLVEMENT SCORES FOR EACH OF THE TEST CONDITIONS

Style of management (A)	Locus of control (B)		
	(b <sub>1</sub> ) Internal (2-6)	(b <sub>2</sub> ) Intermediate (7-10)	(b <sub>3</sub> ) External (11-16)
(a <sub>1</sub> ) Participative (26-32)	$\bar{AB}_{11} = 31.33$	$\bar{AB}_{12} = 26.78$	$\bar{AB}_{13} = 23.11$
(a <sub>2</sub> ) Directive (12-20)	$\bar{AB}_{21} = 29.11$	$\bar{AB}_{22} = 24.80$	$\bar{AB}_{23} = 21.62$

involvement is largely a function of the Internal-External dimension of personality.

#### DISCUSSION

##### *Findings on Satisfaction with Supervision*

Hypothesis 1, which stated that employees who tended toward internality would prefer participative management while those tending toward externality would prefer a more directive management, was supported by the study. The most interesting finding of the study, however, is the apparent strength of the I-E scale in discriminating between subordinates in terms of their responsiveness to differing managerial styles. The strength of the I-E measure in this regard suggests that it has an unrealized potential for use in corporate organizations. A substantial amount of systematic testing remains to be done, however, to confirm its specific applications.

Although further testing and experimentation will be required to confirm the usefulness of the I-E scale as a management tool, the general finding that the personality of subordinates is an important variable in the supervisor-subordinate relationship has important implications. It suggests, for example, that management style alone is insufficient to account for differences in employee satisfaction, and that a broader, more comprehensive theoretical model is needed. Such a model should incorporate and integrate all the major variables that this and other studies have shown to be relevant. One such approach to an integrative theory of supervision is that of Tannenbaum and Schmidt (1958) in which the authors postulate three sets of factors that are of particular importance in determining leadership patterns. These factors are (*a*) forces in the manager, (*b*) forces in the subordinates, and (*c*) forces in the situation.

It is curious that, although Tannenbaum and Schmidt outlined their theory in 1958, relatively little has been done to develop it further or to verify its postulates. One problem with such a theory, of course, is the number of variables which it must encompass; another is its indeterminant character. At best, an interaction theory of supervisory behavior would alert management to the multiplicity of factors which must be considered in determining a supervisory stance. At worst, it would provide no prescriptions for a managerial style that would be appro-

priate for all situations. Despite these limitations, enough evidence is available to suggest that only an interaction theory of the sort proposed by Tannenbaum and Schmidt is adequate for the task.

##### *Findings on Job Involvement*

The findings on the relationship of job involvement to the independent variables are mixed in terms of the study's hypotheses. Hypothesis 2, which states that the job involvement of Internals should be high under participative management, and low under directive management was not supported. Hypothesis 3, which states that the job involvement of Externals should be low, regardless of management style, did find support. Unanticipated by the hypotheses was the finding of an inverse relationship between job involvement and locus of control. That is, Internals tend to score high on the job involvement scale, while Externals tend to score low.

One issue that these findings illuminate is whether work involvement, as defined by the short form of the Lodahl-Kejner scale, is a relatively stable personal characteristic, as suggested by Dubin (1956) and by Lodahl and Kejner (1965); whether it is subject to variation, depending upon situational factors, as suggested by Vroom (1969); or whether it is a concept that is influenced by both personal and situational variables as suggested by Lawler and Hall (1970). Based on this study, the weight of the evidence seems to lie with the "relatively stable personal characteristic" point of view.

In formulating the hypotheses concerning work involvement, the relatively stable aspect of its character was recognized in the prediction that the work involvement of Externals would be low. In the case of Internals, prediction erred in assuming that directive supervision would have a stifling effect on their involvement in their work. The question remains as to how this finding can be reconciled with the concepts that have been employed.

Work involvement, as conceived by Dubin (1956), is intimately bound up in the Protestant Ethic, the moral character of work and a sense of personal responsibility. Anyone who has internalized these traditional values will probably be "work involved" regardless of

the situational context within which he might be employed. If we assume that the Internal in this culture is characterized by a tendency to internalize traditional values about work, then a high correlation between internality and work involvement would seem reasonable, and the hypothesis subordinating work involvement to management style ill-advised.

#### *Major Issues*

Thus far it has been assumed that the explanations of the findings of this study were limited to the theoretical considerations within which they were framed. This assumption may not be valid, and there are alternative explanations of the results that should be recognized. The relationship between the I-E score and work involvement, for example, was explained in terms of the internalization of traditional values concerning work. An alternative interpretation is that this relationship is simply a function of age. Lodahl and Kejner (1965) have reported that there is some evidence that older personnel tend to be more job involved. Since the ages of the subjects in the present study ranged from 21 to 64, one only has to assume that the I-E scale is negatively correlated with age in order to "explain" the study's findings. Although there has been no systematic work done with the I-E scale and age, it is not unreasonable to assume that older people would be more internal than younger ones. This assumption is based on the observation that one of the benefits of increasing age is that, by furnishing additional experience, it provides an opportunity for a more balanced perception of the sources of one's reinforcements. This interpretation would not be inconsistent with Rotter's (1966) concept of locus of control since it is based on learning that, hopefully, is a continuing process.

An alternative explanation for the finding that satisfaction with supervision is a product of the interaction of management style and I-E score is somewhat more complex than the alternative explanation for work involvement. However, if one assumes that (a) older workers tend to be more internal; (b) older workers tend to be more satisfied with their supervision simply because they are glad to have a job; (c) older workers, because of their seniority and experience, tend to

drift into supervisory groups where an easy-going, informal relationship exists; and (d) younger men (regardless of whether they are Internals or Externals) tend to be satisfied or dissatisfied with supervision for a variety of reasons that may have nothing to do with their degree of internality, then it would not be wholly unreasonable to expect the appearance of an interaction effect on the satisfaction scores although none really exists.

In this alternative, as well as in the one concerning job involvement, the age of the respondents is a critical factor. It is unfortunate that the need to protect the anonymity of the subjects of this study precluded the gathering of personal data that would resolve the issue that has been raised.

It is apparent from the foregoing discussion that further work with the I-E scale is needed in order to make it a truly useful research tool. The direction of this future work should include the examination of the I-E scale in relation to certain obvious variables such as age, education, and work experience. Until such work is undertaken, the precise nature of the relationships that may exist remains a matter of speculation.

#### REFERENCES

- DUBIN, R. Industrial workers' worlds: A study of the "central life interests" of industrial workers. *Social Problems*, 1956, 3, 131-142.
- GLESER, G., CRONBACH, L., & RAJARATNAM, N. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 1965, 30, 395-418.
- LAWLER, E. E., III, & HALL, D. T. Relationship of job characteristics to job involvement, satisfaction, and intrinsic motivation. *Journal of Applied Psychology*, 1970, 54, 305-312.
- LIKERT, R. *The human organization*. New York: McGraw-Hill, 1967.
- LODAHL, T., & KEJNER, M. The definition and measure of job involvement. *Journal of Applied Psychology*, 1965, 49, 24-33.
- ROTTER, J. B. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 1966, 80, 1-27.
- TANNENBAUM, R., & SCHMIDT, W. H. How to choose a leadership pattern. *Harvard Business Review*, 1958, 36, 95-101.
- VROOM, V. H. Industrial social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology*. Vol. 5. (2nd ed.) Reading, Mass.: Addison-Wesley, 1969.
- WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

## JOB SATISFACTION AMONG WHITES AND NONWHITES: A CROSS-CULTURAL APPROACH

CHARLES A. O'REILLY, III, AND KARLENE H. ROBERTS<sup>1</sup>

*University of California, Berkeley*

Job satisfaction response patterns were examined for white and nonwhite females across three occupational levels. Three of the most frequently used job satisfaction measures (Job Description Index, GM Faces Scale, Brayfield-Rothe job satisfaction index) were employed. The results of the study suggest that the frame of reference one brings from his culture or subculture influences the way he perceives his job and those facets of it which are satisfying and dissatisfying.

The usual purpose of cross-cultural research is to understand the impact of components of the cultural environment on behavior. These components—generally, beliefs, values, customs, and folkways—are most often compared across rather than within nations. However, as is obvious (Graham & Roberts, 1972; Roberts, 1970), cultural differences can also be found within a single country. When the cultural groups studied are indigenous to the host country, the probabilities are increased that some knowledge of the respective cultures has been accumulated previously by researchers. Thus, a greater understanding of the interactions of variables studied is possible than is often possible in cross-national research.

The condition of nonwhites in the United States approximates that of a subculture and can, therefore, be compared with the predominant white culture. For example, the socialization of a child reared in a black ghetto is recognized as considerably different from that of a middle-class, white, Protestant child. The effects of different socialization processes undoubtedly influence adult behaviors (Clark, 1967). Since socialization of school, work, and social behaviors is different for whites and nonwhites it may be expected that attitudes and behaviors toward work will reflect these subcultural differences. Specifically, this study examines differences in job satisfaction as reported by whites and nonwhites employed in the same work.

Job satisfaction research has demonstrated that different frames of reference affect workers' perceptions of job satisfaction. These influences have been documented for different occupational levels (Armstrong, 1971; Centers & Bugental, 1966; Doll & Gunderson, 1969; England & Stein, 1961), for the male-female dichotomy (Waters & Waters, 1969; Wild, 1970; Williamson & Karras, 1970), for different social levels (Friedlander, 1966), for different educational levels (Klein & Maher, 1966), and even for differences in the characteristics of communities in which the workers live (Hulin, 1966).

The frame of reference the worker brings with him to the job is, then, a determinant of the satisfaction he is likely to derive from it. Hence, should a subculture in the United States provide its members with a different frame of reference from the majority viewpoint, it is anticipated that differences will be reflected in workers' perceptions of job satisfaction. Evidence relevant to the notion that whites and nonwhites have different frames of reference include findings that black children express high levels of vocational aspiration, but low levels of functional striving (Bowerman & Campbell, 1965; Stephensen, 1957). High aspirations and low expectations may easily result in dissatisfaction. Other pertinent findings show that females in the black community enjoy higher academic status than males, both initially and upon termination of their formal education. Black females also exhibit higher educational aspirations than males (Dreger & Miller, 1968). This suggests potential cultural differences between white and black females with respect to the way they

<sup>1</sup> Requests for reprints should be sent to Karlene H. Roberts, School of Business Administration, University of California, Berkeley, California 94720.

view their jobs. Still other evidence shows that nonwhites have lower self-esteem than whites (Haggstrom, 1963). When aggregated, variables like these provide the frame of reference with which a person approaches his job. Therefore, while different frames of reference may contribute to different job satisfaction levels among employees, they may also be related to different patterns of responses to job satisfaction instruments. Such response pattern differences should, then, be interesting to the researcher because they help him understand the etiology of job satisfaction.

The evidence that whites and nonwhites derive satisfaction from different characteristics of their jobs is, unfortunately, inconclusive. A study by Bloom and Barry (1967) tested the Herzberg theory on a sample of black and white workers. However, only 47 of the original sample ( $n = 85$  blacks and 117 whites) returned questionnaires. Furthermore, the total subsample of blacks was unskilled, while 95% of the white subsample was either skilled or semiskilled. Another study looked at satisfaction among underprivileged workers and used a white-nonwhite division of the sample (Champagne & King, 1967). Here, as in the preceding study, differences were found across a cultural dimension. Again, however, there is a question of external validity since the subsamples were not controlled for sex (both men and women were included in the culturally divided subsamples). Champagne and King did not discuss how other factors, such as job level and education, might interact with the cultural split.

#### METHOD

Subjects for this study were drawn from two West Coast hospitals with comparable structure, staffing, and goals. Each subject completed a demographic sheet; the Job Description Index (JDI; Smith, Kendall, & Hulin, 1969) which measures satisfaction with five aspects of the job (the work itself, WJDI; satisfaction with coworkers, COJDI; satisfaction with supervision, SUJDI; satisfaction with pay, PAJDI; and satisfaction with opportunities for promotion, PRJDI); the GM Faces Scale (Kunin, 1955), an overall measure of satisfaction using a projective rather than a descriptive technique; and the Brayfield-Rothe index (Brayfield & Rothe, 1951), an 18-item measure of overall satisfaction.

Response rate for the entire sample ( $n = 495$ ) was 85%. From this total, a matched sample was ob-

tained ( $n = 139$ ) by using all nonwhite respondents and matching them with white counterparts. The match resulted in two subsamples, one white and one nonwhite. The two subsamples consisted of respondents matched by occupational level and education. All subjects were female, full-time employees, stratified into three occupational groups (registered nurses and supervisors—RNs— $n = 38$ ; licensed vocational nurses—LVNs—and technicians,  $n = 51$ ; aides and clerical personnel,  $n = 50$ ). A comparison of differences in means and standard deviations across 60 demographic variables for the two cultural subsamples revealed a highly homogeneous population. The two subsamples were then compared on job satisfaction measures. The total white and nonwhite samples were first compared ( $n = 69$  whites and 70 nonwhites). This was followed by analyses of the white-nonwhite splits at each of the three occupational strata indicated above.

Analyses included *t* tests and omega squared tests of the differences in means between whites and non-

TABLE 1

*t* TESTS AND TESTS FOR STRENGTH OF RELATIONSHIPS  
(OMEGA SQUARED) BETWEEN WHITES  
AND NONWHITES

Sample and scale	N	<i>t</i> statistic	Omega squared
Unstratified sample			
GM Faces	138	4.84**	.14
WJDI	135	3.92**	.10
COJDI	135	2.37*	.03
SUJDI	135	3.13**	.06
PAJDI	135	4.32**	.12
PRJDI	135	-.24	—
RNs and supervisors			
GM Faces	38	3.38**	.22
WJDI	38	1.65	.04
COJDI	38	.59	—
SUJDI	38	.30	—
PAJDI	38	1.50	.03
PRJDI	38	.36	—
LVNs and technicians			
GM Faces	50	1.78	.04
WJDI	49	2.12*	.07
COJDI	49	.52	—
SUJDI	49	1.46	.03
PAJDI	49	2.51*	.09
PRJDI	49	-1.02	—
Aides and clerical			
GM Faces	50	3.40**	.17
WJDI	48	2.99**	.15
COJDI	48	2.82**	.13
SUJDI	48	3.88**	.23
PAJDI	48	4.48**	.29
PRJDI	48	.05	—

Note. Results of *t* tests report means of the white samples significantly higher than nonwhites.

\* $p < .05$ .

\*\* $p < .01$ .

TABLE 2  
CORRELATIONS BETWEEN GM FACES AND JDI SCALES

GM Faces	JDI scales				
	WJDI	COJDI	SUJDI	PAJDI	PRJDI
Unstratified sample ( $N = 139$ )					
Whites ( $n = 69$ )	.57**	.17	.40**	.36**	.36**
Nonwhites ( $n = 70$ )	.33**	.32**	.22	.24*	.35**
RNs and supervisors ( $N = 38$ )					
Whites ( $n = 19$ )	.57*	-.04	.43	.03	.68**
Nonwhites ( $n = 19$ )	.32	.14	.39	.46	.12
LVNs and technicians ( $N = 51$ )					
Whites ( $n = 25$ )	.73**	.32	.46*	.28	.28
Nonwhites ( $n = 26$ )	.57**	.37	.38*	.31	.47*
Aides and clerical ( $N = 50$ )					
Whites ( $n = 25$ )	.32	.09	.17	.55*	.20
Nonwhites ( $n = 25$ )	.20	.37	.06	.09	.46*

Note. For the unstratified sample,  $r = .30$  for  $p < .01$ ,  $r = .23$  for  $p < .05$ ,  $df = 69$ ; for RNs and supervisors,  $r = .58$  for  $p < .01$ ,  $r = .46$  for  $p < .05$ ,  $df = 17$ ; for LVNs and technicians, and aides and clerical,  $r = .49$  for  $p < .01$ ,  $r = .38$  for  $p < .05$ ,  $df = 25$ .

\* $p < .05$ .

\*\* $p < .01$ .

whites for scores on the three satisfaction instruments. To uncover differences in response patterns across the two cultural subsamples, intercorrelation matrices for the satisfaction instruments were compared. Particular attention was paid to intercorrelations among the JDI and GM Faces Scale. The intent was to use the GM Faces Scale, an indicator of overall satisfaction, as a general measure and then to determine the importance of various JDI scales as components of overall satisfaction. Principal components factor analyses were also run for the stratified subsamples. However, due to the small sample size these results were too speculative to report.

### RESULTS

Results of the  $t$  tests and omega squared tests for the entire unstratified sample ( $n = 139$ ) are reported in Table 1.

As indicated by the GM Faces Scale and four JDI scales for the unstratified sample, whites were significantly more satisfied with their jobs than were nonwhites. This is corroborated by intercorrelations of Brayfield-Rothe items with both the GM Faces and JDI scales (not shown).

Results of the  $t$  tests for each of the three occupational strata (RNs, LVNs, and aides) are also given in Table 1. As is evident, the strongest differences were in the highest stratum (RNs) and the lowest (aides). Again, whites were considerably more satisfied with their jobs than were nonwhites. Although not

shown, at each stratum a number of significant Brayfield-Rothe items also confirmed the higher satisfaction levels of the whites.

Consideration of the overall comparisons of the unstratified sample ignores differences in the response patterns for subcultural groups at different occupational levels. Thus, intercorrelation matrices of the GM Faces and JDI scales were examined for the white and nonwhite groups, both for the unstratified sample and the three occupational levels. The intercorrelations of Brayfield-Rothe items with the other scales are not shown here because they do not substantially add to the information presented.

For the RNs and supervisors, examination of Table 2 shows that for whites, the GM Faces Scale, the overall measure of satisfaction, correlates most highly with the JDI scale representing satisfaction with promotion (PRJDI). At the same time, the GM Faces Scale correlates negatively with the JDI score representing the importance of co-workers (COJDI). Intercorrelations for nonwhite RNs and supervisors reveal they have a different perspective. In this case, PRJDI is not significantly correlated with GM Faces. If anything is important for nonwhites it seems to be PAJDI, satisfaction with pay.

Referring to Table 1, at the second occupa-

TABLE 3

CORRELATIONS BETWEEN THE WJDI SCALE  
AND OTHER JDI SCALES

WJDI	Other JDI scales			
	COJDI	SUJDI	PAJDI	PRJDI
Aides and clerical (N = 50)				
Whites (n = 25)	.55**	.52**	.49**	.46*
Nonwhites (n = 25)	.72**	.76**	.13	.30

\*  $p < .05, r = .38, df = 25.$

\*\*  $p < .01, r = .49, df = 25.$

tional stratum, LVNs and technicians, whites were more satisfied with their jobs than were nonwhites. Table 2 presents the intercorrelations between JDI and GM Faces scales for this job level. There are no major differences in patterns for the two subsamples. However, while the results of the factor analyses must be treated as highly speculative, the data from the two subsamples at this level show both whites and nonwhites excluding PRJDI from the large general factor loading. White LVNs and technicians included PRJDI in a well-defined second factor. Nonwhites completely excluded consideration of promotion from the principal loadings.

At the lowest job level, aides and clerical personnel, Table 1 again shows whites as considerably more satisfied than nonwhites. The intercorrelations of GM Faces and the JDI scales (Table 2), however, reveal no noteworthy differences in response patterns for the two groups. A closer look at the data (Table 3) indicates intercorrelations of WJDI (satisfaction with work, a possible surrogate of overall satisfaction) with the other four JDI scales to be different for whites and nonwhites. The WJDI scale is significantly correlated with all JDI scales for whites, but only with COJDI and SUJDI for nonwhites. The COJDI and SUJDI scales are thought to represent extrinsic work factors. Factor analyses, though speculative, support this trend for nonwhite concern with extrinsic job factors and white concern with both extrinsic and intrinsic factors.

## DISCUSSION

Without theoretical notions to explain culture and to predict its influence on other vari-

ables, it is difficult to make sense of cross-cultural or subcultural comparisons. Yet, it is extremely important that researchers seek the etiology of the differences they find. In the case presented here, because American researchers previously have studied American subcultures, our understanding of the relevant variables in the two cultures should help suggest explanations for the differences in the way people respond to their jobs. This is a step better than merely identifying differences in traits between two cultures.

Many possible influences on job satisfaction are controlled here because of the similarity of the white and nonwhite samples. Findings are based on three of the most widely used and best researched job satisfaction instruments. Thus, the probability of erroneous findings resulting from bias in any single instrument is reduced and convergent validity obtained. However, since the data reported are primarily correlational, no case is made for causation. Rather, the data are suggestive of some possible influences on job satisfaction which should be more carefully examined across subcultural groupings.

Closer examination of the white and nonwhite subsamples at the three occupational levels was necessary to arrive at some highly tentative suggestions about possible cultural variations in satisfaction. White RNs and supervisors, for example, not only were more satisfied than nonwhites but associated to a greater degree overall job satisfaction (GM Faces) and satisfaction with promotion (PRJDI). Differences in satisfaction for LVNs and technicians were not as great. This, of course, might come from an initial expectation that promotion is impossible to obtain. At the lowest occupational level studied, aides and clerical personnel, nonwhites were again significantly less satisfied than their white counterparts and there is minor evidence that nonwhites were concerned with the social factors of their jobs and whites with these as well as pay and promotional opportunities. Future research should seek such differences in response patterns across subcultural groups and clarify the reasons for them.

Data analyses here generally illustrate the existence of job satisfaction differences across

a cultural dichotomy. This is borne out by differences in relative levels of satisfaction as illustrated by the three instruments used, and by differences in response patterns between the white and nonwhite groups at three occupational levels. These differences should be verified by future research. Explanations of them are purely speculative. We assume that whites and nonwhites approach their jobs with different frames of reference which can be identified and which are related to their job satisfaction. Some empirical evidence supports this notion and our findings. Unfortunately, the correlational evidence presented here and the small *Ns* across the three occupational strata render our results only suggestive with further research required to adequately explicate the differences in levels of job satisfaction and patterns of response for the subcultural groups.

## REFERENCES

- ARMSTRONG, T. B. Job content and context factors related to satisfaction for different occupational levels. *Journal of Applied Psychology*, 1971, 55, 57-65.
- BLOOM, R., & BARRY, J. Determinants of work attitudes among Negroes. *Journal of Applied Psychology*, 1967, 51, 287-292.
- BOWERMAN, C., & CAMPBELL, E. Aspiration of southern youth: A look at racial comparisons. *Trans-Action*, 1965, 2, 24.
- BRAYFIELD, A., & ROTHE, H. An index of job satisfaction. *Journal of Applied Psychology*, 1951, 35, 307-311.
- CENTERS, R., & BUGENTAL, D. Intrinsic and extrinsic job motivations among different aspects of the working population. *Journal of Applied Psychology*, 1966, 50, 187-196.
- CHAMPAGNE, J., & KING, D. Job satisfaction factors among underprivileged workers. *Personnel and Guidance Journal*, 1967, 45, 429-434.
- CLARK, K. Explosion in the ghetto. *Psychology Today*, 1967, 1, 30-39.
- DOLL, R., & GUNDERSON, E. Occupational group as a moderator of the job satisfaction-job performance relationship. *Journal of Applied Psychology*, 1969, 53, 359-361.
- DREGER, R., & MILLER, K. Comparative psychological studies of Negroes and whites: 1959-1965. *Psychological Bulletin*, 1968, 70, 1-32.
- ENGLAND, G., & STEIN, C. The occupational reference group: A neglected concept in employee attitude studies. *Personnel Psychology*, 1961, 14, 299-308.
- FRIEDLANDER, F. Importance of work versus non-work among socially and occupationally stratified groups. *Journal of Applied Psychology*, 1966, 50, 430-439.
- GRAHAM, W., & ROBERT, K. *Comparative studies in organizational behavior*. New York: Holt, Rinehart & Winston, 1972.
- HAGGSTROM, W. Self-esteem and other characteristics of residentially desegregated Negroes. *Dissertation Abstracts*, 1963, 23, 3007.
- HULIN, C. Effects of community characteristics on measures of job satisfaction. *Journal of Applied Psychology*, 1966, 50, 185-192.
- KLEIN, S., & MAIER, J. Educational level and satisfaction with pay. *Personnel Psychology*, 1966, 19, 195-208.
- KUNIN, T. The construction of a new type of attitude measure. *Personnel Psychology*, 1955, 8, 65-77.
- ROBERTS, K. On looking at an elephant: An evaluation of cross-cultural research related to organizations. *Psychological Bulletin*, 1970, 74, 327-350.
- SMITH, P., KENDALL, L., & HULIN, C. *The measurement of satisfaction in work and retirement*. Chicago: Rand McNally, 1969.
- STEPHENSON, R. Mobility orientation and stratification of 1,000 ninth graders. *American Sociological Review*, 1957, 22, 204-212.
- WATERS, L., & WATERS, C. Correlates of job satisfaction and job dissatisfaction among female clerical workers. *Journal of Applied Psychology*, 1969, 53, 388-391.
- WILD, R. Job needs, job satisfaction, and job behavior of women manual workers. *Journal of Applied Psychology*, 1970, 54, 157-162.
- WILLIAMSON, T., & KARRAS, E. Job satisfaction variables among female clerical workers. *Journal of Applied Psychology*, 1970, 54, 343-348.

(Received October 26, 1971)

## THE NATURE OF BIAS IN OFFICIAL ACCIDENT AND VIOLATION RECORDS<sup>1</sup>

FREDERICK L. McGUIRE<sup>2</sup>

*University of California, Irvine*

Since many studies in accident research derive criteria from official records, the existence of systematic biases in these files could have profound implications. This study demonstrates that accident and citation frequency are grossly underrecorded and that biases exist by sex, age, occupation, and race.

In the field of accident research, it is common practice to utilize official motor vehicle records as sources of criteria (e.g., accidents and citations). It is generally known that these events are underreported, but should these records also contain systematic biases it would have profound implications for those studies on which such conclusions are based. This study was designed to identify some of these biases by comparing information obtained in confidential interview with those obtained from official motor vehicle records.

### METHOD

Approximately 6,000 successful license applicants at the headquarters of the Highway Patrol in Jackson, Mississippi, had completed a biographical questionnaire. On the 2-year anniversary of licensure an attempt was made to interview each subject and to obtain his/her accident and citation record. Eventually, 2,797 persons were interviewed, primarily by telephone. For purposes of comparing the data obtained from the interview with that in the official records, only the first 500 cases were used.

In this study an accident is defined as any collision taking place while the subject was operating the motor vehicle, regardless of damage or fault. Instances in which the subject was legally motionless, such as when parked or at a stop light, are excluded. Citations are defined as a "ticket" actually received for a moving

<sup>1</sup> The project upon which this publication is based was performed under support of Grant UI 00043 (formerly AC 00313) and administered by the Bureau of Community Environmental Management. These results do not necessarily reflect policy or recommendations advanced by the Department of Health, Education, and Welfare.

Appreciation is expressed to R. C. Kersh and John Plag for their help in gathering and analyzing these data and to Myra Ellington for her assistance in preparing the manuscript.

<sup>2</sup> Requests for reprints should be sent to Frederick L. McGuire, Department of Psychiatry and Human Behavior, University of California, Irvine, California 92664.

violation; parking tickets and equipment violations, for example, are not included. Mississippi law provided that an accident was reportable if \$50 of damage or personal injury was incurred.

### RESULTS AND DISCUSSION

Of the 110 reportable accidents described during the interview as taking place among these 500 subjects, 74 or 67% were, in fact, said to have been actually reported in writing, as required by law, but only 42% appeared in the files. In the larger sample of 2,797 cases, the drivers stated that 75% of their reportable accidents had been reported (Table 1). Although other factors could account for some of this discrepancy, by the subject's own admission it is evident that most of the "missing" reportables are simply not reported. When all accidents admitted to during interview of these 500 subjects (reportables plus nonreportables) were added together, it was found that only 25% were in the records (57 of 230).

In terms of the number of drivers (as opposed to the number of accidents) not properly included in the state records, it was found that of these 500 subjects, 53 or 11% were officially listed as having an accident, although 110 or 22% admitted to having reportable accidents. Thirty-eight percent admitted to either a reportable or nonreportable accident during the 2 years.<sup>3</sup>

Since a serious accident is more likely to be reported, the entire sample of 2,797 subjects was examined according to frequency of claimed reporting and extent of damage or injury. For reportable accidents under \$100

<sup>3</sup> In the majority of cases, each interview event was determined to be the same event noted in the record. If this was not clear the researcher used dates and descriptions to arrive at a considered opinion.

it appears that there is about 47% chance of the event being reported, according to interview, but when the damage is over \$100 the odds rise sharply (Table 1).

In the case of personal injuries the percentage of accidents said to have been reported rises from 72% with no injury to 85% with minor injuries, 91% in cases of hospitalization, and 100% when a fatality occurred. It is apparent, then, that reportable accidents tend to be reported largely as a function of damage and/or injury.

In order to determine to what extent other biases exist in these records, a comparison between the two sources of data was made according to sex, occupational category, age at time of licensure, and race.

Bias was defined as the percentage of interview events that appeared in the official motor vehicle records for each subject (i.e., number of events in official records divided by number of interview events). The correlation ( $r$ ) of these percentage scores with the segmented variables of age, sex, etc., provided a measure of the degree of association between overall bias in the official records and each characteristic. All correlations were corrected for curvilinearity.

*Age.* Table 2 shows that each age group is significantly underrepresented in the state records. Although a trend in Table 2 appeared to exist in favor of younger drivers having a higher percentage recorded, the correlation between age and level of reporting was not significant. The fact that accident records do

TABLE 1  
NUMBER OF REPORTABLE ACCIDENTS ADMITTED IN  
INTERVIEW VERSUS NUMBER SAID TO HAVE  
BEEN OFFICIALLY REPORTED—BY  
AMOUNT OF DAMAGE

Damage	Total number of reportable accidents by interview	Said reported	
		n	%
Under \$50	17	8	47
\$50-\$100	164	87	53
\$101-\$500	257	219	85
Over \$500	45	41	91
Demolished	52	45	87
Total	535	400	75

Note.  $N = 2,797$ .

not reflect an age bias is a finding not predicted by some authors (Klein, 1966).

*Occupation.* Table 3 indicates an overall bias according to occupational category. As noted by the  $t$  test results within categories, each group contributes significantly to this bias, with the semiprofessional and professional category contributing the least.

*Race.* Both blacks and whites are significantly underrepresented in the records (Table 4). Although it appears that black drivers are more likely to have an accident recorded than are white drivers, the correlation between proportion of accidents in Highway Patrol records and race was not significant.

*Sex.* Although both males and females are significantly underrepresented in the state records (Table 5), the female group had only

TABLE 2  
REPORTABLE ACCIDENTS OBTAINED BY INTERVIEW VERSUS THOSE FOUND IN RECORDS  
OF HIGHWAY PATROL—BY AGE

Age	No. subjects with accidents <sup>a</sup>	No. accidents by interview			No. accidents in Highway Patrol records				<i>t</i> ratio, within category
		n	M	SD	n	M	SD	%	
15-19	59	62	1.05	.43	29	.49	.53	47	5.92*
20-29	15	17	1.13	.34	7	.47	.50	41	4.19*
30+	27	31	1.15	.45	10	.37	.48	32	5.05*
Total	101	110	1.09	.43	46	.46	.52	42	8.72*

Note.  $N = 500$ . Correlation between proportion of interview accidents in Highway Patrol records and age:  $r = .08$ , ns (see text for method of computation).

<sup>a</sup> Of 500 subjects studied, 101 incurred 110 accidents.

\*  $p < .001$ .

TABLE 3

## REPORTABLE ACCIDENTS OBTAINED BY INTERVIEW VERSUS THOSE FOUND IN RECORDS OF HIGHWAY PATROL—BY OCCUPATION

Occupational category	No. subjects with accidents <sup>a</sup>	No. accidents by interview			No. accidents in Highway Patrol records				<i>t</i> ratio, within category
		<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	%	
Student	57	60	1.05	.44	28	.49	.53	47	5.79**
Housewife	9	13	1.30	.46	1	.10	.30	8	6.00**
Unskilled									
Semiskilled									
Skilled	23	25	1.09	.28	13	.57	.50	52	4.22**
Semiprofessional									
Professional	11	12	1.09	.52	4	.36	.48	33	2.67*
Total	101	110	1.09	.43	46	.46	.52	42	8.72**

Note. *N* = 500. Correlation between proportion of interview accidents in Highway Patrol records and occupational category:  $r = .25$ ,  $p < .02$  (see text for method of computation).

<sup>a</sup> Of 500 subjects studied, 101 incurred 110 accidents.

\*  $p < .05$ .

\*\*  $p < .001$ .

26% of their reportable accidents recorded as compared with 53% for the males. This is a significant difference reflected in the correlation of .22 ( $p < .05$ ) between proportion of accidents in Highway Patrol records and sex. This difference is partially explained by the fact that women said they reported 55% of their reportable accidents, as opposed to 73% for men. Beyond this, other sex-linked hypotheses must be entertained.

To summarize, there were no age or race biases; any one age group or race was as likely to be reported as any other age group or race. The occupational and sex categories were correlated significantly with the proportion

of accidents in Highway Patrol records. For the occupational category, the semiprofessional and professional group contributed less to this correlation than did any of the other occupational categories. For sex, females were reported to a significantly lesser degree than were males.

An analysis was also performed on the frequency of citations reported.<sup>4</sup> An overall bias

<sup>4</sup> The tables for citations are not presented in order to conserve space; the method of analysis was identical to that with accidents. When analyzed according to "with accident" and "without accident," no differences were found; findings, therefore, relate to combined citations.

TABLE 4

## REPORTABLE ACCIDENTS OBTAINED BY INTERVIEW VERSUS THOSE FOUND IN RECORDS OF HIGHWAY PATROL—BY RACE

Race	No. subjects with accidents <sup>a</sup>	No. accidents by interview			No. accidents in Highway Patrol records				<i>t</i> ratio, within category
		<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	%	
White	91	100	1.10	.45	40	.44	.52	40	8.41**
Black	10	10	1.00	.00	6	.60	.49	60	2.45*
Total	101	110	1.09	.43	46	.46	.52	42	8.72**

Note. *N* = 500. Correlation between proportion of interview accidents in Highway Patrol records and race,  $r = .11$ ,  $ns$  (see text for method of computation).

<sup>a</sup> Of 500 subjects studied, 101 incurred 110 accidents.

\*  $p < .05$ .

\*\*  $p < .001$ .

TABLE 5  
REPORTABLE ACCIDENTS OBTAINED BY INTERVIEW VERSUS THOSE FOUND IN RECORDS  
OF HIGHWAY PATROL--BY SEX

Sex	No. subjects with accidents <sup>a</sup>	No. accidents by interview			No. accidents in Highway Patrol records				<i>t</i> ratio, within category
		<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	%	
Male	61	64	1.05	.38	34	.56	.53	53	5.72*
Female	40	46	1.15	.48	12	.30	.46	26	6.99*
Total	101	110	1.09	.43	46	.46	.52	42	8.72*

Note. *N* = 500. Correlation between proportion of interview accidents in Highway Patrol records and sex:  $r = .22$ ,  $p < .05$  (see text for method of computation).

<sup>a</sup> Of 500 subjects studied, 101 incurred 110 accidents.

\*  $p < .001$ .

in reporting levels was found for age ( $r = .27$ ,  $p = .001$ ), race ( $r = .16$ ,  $p = .05$ ), and sex ( $r = .26$ ,  $p = .001$ ). Whites and females had significantly fewer citations recorded, but in terms of age the results were nonlinear; the 15-year-olds and the 20-29-year-olds were significantly underrecorded; those aged 16-19, 30-39, and 40 and above were not. No overall bias in citation reporting was noted for occupational category, but the semiprofessional, professional group did have 100% of their citations in the records.

Thus, with regard to citations, there were significant age and race biases that did not

exist in the case of reported accidents. There was a sex bias for both, and an occupational bias for accidents only.

#### COMPARISON WITH OTHER STATES

It may be questioned if the Mississippi data are generalizable to other states. Table 6 compares those gathered in Mississippi with one group from Illinois and three from California. Ross (1966) interviewed 36 accident-involved drivers, recorded the number of accidents and citations admitted to by the subjects, and compared these with official Illinois records.

TABLE 6  
COMPARISON OF ACCIDENTS AND CITATIONS OBTAINED BY INTERVIEW AND OTHER SOURCES  
VERSUS THOSE CONTAINED IN STATE RECORDS

State	Group	Reportable accidents			Citations		
		No. by interview	No. in state record	% in state record	No. by interview	No. in state record	% in state record
California <sup>a</sup>	120 high school seniors	46	35	76	56	57	100 <sup>b</sup>
Mississippi <sup>b</sup>	500 license applicants	110	46	42	196	150	77
Mississippi <sup>c</sup>	500 license applicants	73	39	53	—	—	—
Illinois <sup>d</sup>	36 accident involved drivers	25	9	36	25	18	72
California <sup>e</sup>	438 state employees	—	—	65 <sup>d</sup>	—	—	—
California <sup>f</sup>	3,842 license applicants	—	—	66 <sup>e</sup>	—	—	—

<sup>a</sup> About one-quarter of follow-up period under reporting level of \$200 personal injury.

<sup>b</sup> Reporting level over \$50 personal injury.

<sup>c</sup> Reporting level over \$100 personal injury.

<sup>d</sup> Comparison of employee's report to supervisor versus Department of Motor Vehicle records. Includes only collisions with \$100 damage to most severely damaged vehicle (Smith, 1966).

<sup>e</sup> Comparison of Department of Motor Vehicle records versus insurance records. No distinction made between reportable and nonreportable accidents (Burg, 1967).

<sup>f</sup> Even though one more citation was contained in the state record than was obtained by interview, 100% was used as an upper limit.

(Illinois law defines a reportable accident as \$100 damage and/or personal injury.)

Data from the first California group were gathered from an interviewed sample of 120 high school seniors for whom it was determined that at least 46 reportable accidents occurred (McGuire & Kersh, 1969). (Until January 1, 1968, the California criteria for reporting was \$100 damage and/or personal injury; since then, the criteria is \$200 damage and/or personal injury.)

The Mississippi data were subdivided into those also meeting the \$100 minimum, in order to be more comparable to California and Illinois data.

As noted in Table 6, a number of accidents are not included in each state file: California records show only 76% (for the teenagers), Mississippi 42%, and Illinois 36%; when corrected for the \$100 minimum, the Mississippi level increased to 53%. Since reporting levels are related to severity, and the California sample includes several months of exposure under the minimal reporting level of \$200, the higher level for the state may be inflated. In terms of citations, California records contained all those citations obtained by interview, while Mississippi and Illinois records listed only 77% and 72%, respectively.

The fact that California recorded 100% of all known citations speaks well for the quality of its reporting and retrieval system. The omission of 24% of known accidents is probably due mostly to driver secrecy.

In order to compare two of the states in terms of age and sex, a sample was extracted from these 500 Mississippi drivers of the same age and sex as a sample from the 120 high school seniors from California. It was found that the male teenagers from Mississippi ( $n = 122$ ) had 58% of their admitted collisions listed in the official records versus 79% for the California males ( $n = 90$ ), a difference of 21%. The Mississippi females ( $n = 73$ ) had only 29% recorded, while their California counterparts ( $n = 30$ ) had 63% listed, a difference of 34%. This suggests that both states show a sex bias in their level of reporting, but that it is more pronounced in Mississippi.

The second California group resulted from a study conducted on a sample of 438 accidents incurred during 1963 by state-owned vehicles

driven by adult employees of the California Division of Highways (Smith, 1966). Since these employees were under strict orders to report all accidents to their own division no matter how minor, this provided a means of comparing such reports with those located in the official files of the Department of Motor Vehicles. Because the latter were received from local and state law enforcement agencies, and at least one such agency, the California Highway Patrol, attempted to report all accidents brought to their attention, these data apparently include a mixture of legally reportable and nonreportable collisions.

In general, Smith found that about 49% of all accidents were recorded, with 65% of those showing \$100 damage or more to the most severely damaged vehicle; the Mississippi level was 42% for all reportable accidents and 53% when damage to the subject's car was \$100 or more and/or a personal injury was included. The California group had 93% of all injury accidents recorded in the state records, while Mississippi drivers claimed to have reported 88%; both groups indicated 100% of all fatalities recorded.

In the third California study, Burg (1967) utilized a combination of Department of Motor Vehicle and insurance records covering a 3-year period, using subjects taken from customer lines at Department of Motor Vehicle offices. Based on 3,842 subjects and a total of 1,316 accidents, he found only 66% of the insurance-reported collisions in the Department of Motor Vehicle files. About 27% of Burg's sample had been accident-involved within 3 years, a figure well above that 15% of the general driving population likely to be recorded (Coppin, 1965), and due in part to the fact that both reportable and nonreportable accidents were included for this group (A. Burg, personal communication, June, 1970).

The difference between reporting levels for the high school seniors and the two other California samples is probably due in large part to the fact that part of the time the younger group operated under a higher minimum for reporting, and only those accidents incurring at least \$200 damage had to be reported. As previously noted, a higher level of reporting exists for the more serious collisions.

In a separate study of Illinois motor vehicle records, Michalski (1965) states that only 33% of an estimated 320,672 accidents during 1958 could be accounted for by official records. This corresponds closely with the 36% (see Table 6) noted by Ross (1966) for the same state and is not very different from the 42% derived from Mississippi records.

### CONCLUSIONS

It seems appropriate to conclude that most states have difficulty in maintaining complete records, which make interstate comparisons difficult. Not only do such records underrepresent actual frequency, but they probably contain definite sex, age, and occupational biases. As with all such data, findings of the present study cannot be generalized without care, but they do underline the fact that the nature of biases existing in official records should first be established before they are used for research or used to form the basis for action programs.

### REFERENCES

- BURG, A. *The relationship between vision test scores and driving record: General findings.* (Report No. 67-24) Los Angeles: University of California, Institute of Traffic and Transportation Engineering, June 1967.
- COPPIN, R. S. *The 1964 California driver record study. Accidents, traffic citations, and negligent operation counts by sex.* (Report No. 20, Part 2) Sacramento, Calif.: State of California, Department of Motor Vehicles, March 1965.
- KLEIN, D. A reappraisal of the violation and accident data on teen-aged drivers. *Traffic Quarterly*, 1966, 20, 502-510.
- MCGUIRE, F. L., & KERSH, R. C. A study of history, philosophy, research methodology, and effectiveness in the field of driver education. *University of California Publications in Education*, 1969, 19, 57-60.
- MICHALSKI, C. S. *Traffic accident data project.* Chicago, Ill.: National Safety Council, 1965.
- ROSS, H. L. Driving records of accident-involved drivers. *Traffic Safety Research Review*, 1966, 10, 22-25.
- SMITH, R. N. The reporting level of California State Highway accidents. *Traffic Engineering*, 1966, 29, 20-25.

(Received December 13, 1971)

## PREDICTION OF ACCIDENTS IN A STANDARDIZED HOME ENVIRONMENT<sup>1</sup>

JOAN S. GUILFORD<sup>2</sup>

*Sheridan Psychological Services, Beverly Hills, California*

A kitchen laboratory was used for the study of accidents incurred by 226 female subjects who performed standardized household tasks under observation. Four years of driving records were obtained for a subsample of 178 subjects possessing licenses. Kitchen criteria were classified as property damage accidents and personal injury accidents, summed to provide total kitchen accidents. Near accidents constituted the fourth kitchen criterion. Significant ( $p < .05$ ) correlations were found between automobile accidents, automobile violations, and kitchen criteria. A number of demographic, attitudinal, physiological, and cognitive predictors correlated significantly ( $p < .05$ ) with both total kitchen accidents and automobile accidents. Environmental control of exposure to hazards made it possible to extend accident criteria to include other behaviors.

The relative lack of success in finding predictors that are consistently and meaningfully related to accident indexes has provided for decades of frustration and generated volumes of controversy. The concept of "accident proneness" (Farmer & Chambers, 1926), or human variability with respect to a characteristic that might be termed accident susceptibility, has fallen into disrepute largely because of methodological problems in demonstrating its existence.

Failure to predict accidents is often less a function of the predictors than of the criteria. Reportable accidents occur so rarely to so few that the time required to establish a reliable criterion weighs heavily against the accident research investigator. Furthermore, the population under study is constantly shifting by virtue of the fact that each fatal or debilitating accident removes the victim from the population at risk, at least temporarily. Also, if public records are used as the basis for determining accident rates, the data derived are

notoriously undependable (McGuire, 1971). Finally, and most important, the occurrence of an accident is a function of exposure to risk, and degree of such exposure is extremely difficult to determine.

Despite the controversy as to whether there is such a thing as accident susceptibility based on human characteristics, Whitlock, Clouse, and Spencer (1963) astutely point out:

Since the psychologist is concerned with the study of behavior, for an injury to qualify as an object of inquiry for the psychologist, the injury must be shown to result from human behavior. The object of interest is the "accident behavior" . . . the injuries the psychologist can hope to predict are only those which do result from accident behaviors [p. 35].

Because accident incidents array themselves in a form best fit by the Poisson distribution, the determination of predictable variance has been based on departure from this chance model. Estimates of the percentage of predictable variance to be found in accident records have ranged from 3% or 4% (Forbes, 1957) to as high as 62% (Thorndike, 1951). Mintz and Blum (1949), in their review, found it to vary from 20% to 40%. In general, the figure is somewhere around 25%, and as Cobb (1940) has shown, at this level, a perfect test of accident proneness cannot correlate better than .44 with an accident criterion.

Methods used to overcome this problem have included the use of minor accidents (e.g., Keehn, 1959; Kunkle, 1946; Newbold, 1926; Wong & Hobbs, 1949) and near accidents (e.g.,

<sup>1</sup> This investigation was supported in whole by the U.S. Public Health Service, Research Grant AC00141 from the Division of Accident Prevention to the American Institutes for Research.

Computing assistance was obtained from the Health Sciences Computing Facility, University of California, Los Angeles, sponsored by National Institutes of Health Grant FR 3.

<sup>2</sup> Requests for reprints should be sent to Joan S. Guilford, Sheridan Psychological Services, P.O. Box 6101, Orange, California 92667.

This research project was conducted while the author was employed at the American Institutes for Research and is reported more fully in Guilford (1965).

Vasilas, Fitzpatrick, DuBois, & Youtz, 1952) to provide more reliable criteria. Errors have also been found to relate to accidents (e.g., Eno Foundation, 1948; Kraft & Forbes, 1944; Ruch & Wilson, 1948). Brody (1962) suggested the possibility of defining accidents not so much in terms of their outcome as in terms of certain "unsafe practices" that could lead to an accident. Whitlock et al. (1963) found specimens of "unsafe performance" over a 1-year period to have a reliability of .93, while over a 4-year period the reliability of injury data for the same sample was only .52. The correlation between injuries and unsafe behaviors was .44, an estimate restricted by the lack of reliability of the injury criterion.

The purpose of this research was to explore in a controlled laboratory setting the possibility of redefining the term "accident" in an effort to develop what might be called "accident behavior criteria" and then to relate human characteristics to these criteria.

Specifically, the inspiration for this approach came from suggestions by Arbous and Kerrich (1951) who advocated extending the accident criterion to include minor accidents or injuries, near accidents, and accident-related behaviors such as errors or slips. They also advocated the structuring of an environment equated for all subjects in which the study of errors might be made in a series of test situations. With environmental variability eliminated through laboratory control and standardization of experimental tasks, it could logically be assumed that any remaining nonchance variance in accidents would be attributable to human variability and, further, predictable from the characteristics of those incurring such accidents or exhibiting such accident behaviors.

## METHOD

### Laboratory Design

The laboratory setting in which data were collected was a mobile van into which were built a simulated home kitchen, two observation rooms with one-way screens, and a testing room. Every item in the kitchen, down to the last utensil, was placed in the same location and position for each subject. Equipment was modern, new, and with the exception of those items altered to increase hazards, in excellent working order. The kitchen was 8 × 11 feet before installation of stove, refrigerator, sink, counters, and kitchen cabinets. This type of laboratory was chosen because (a) home

accidents are responsible for more disabling injuries than any other single class of accident, (b) within the home, the kitchen is the most frequent site of the accident for women, and (c) it was desirable to structure the experimental situation to appear to be as natural as possible for subjects who were women.

### Tasks

A series of kitchen tasks were standardized in such a way that every subject was required to perform the same operations. Operations were selected to maximize accident potential and, at the same time, also to maximize the opportunity for accident-avoiding behavior. Tasks involved baking cupcakes; hardboiling four eggs; preparing a bacon, lettuce, and tomato sandwich on toast; preparing cole slaw; serving two lunches; washing all dishes and utensils; washing nylons and a blouse; ironing the blouse; putting everything away; and cleaning the kitchen. The average time required for these tasks was 2 hours.

### Subjects

The subjects were 226 women who responded to radio and newspaper solicitations to participate in an evaluation of the kitchen and its equipment. The actual purpose of the experiment was not revealed. All subjects were screened to ascertain that they had (a) at least 3 years' experience in homemaking and (b) used a gas stove for at least the past 3 years and could, consequently, use the laboratory stove without difficulty. The subjects came from all parts of Los Angeles and its environs. The laboratory was moved about in order to obtain a wide geographic sampling.

With respect to demographic characteristics, the mean age of the subjects was 37; 91% were married; median years of education was 12.9 for the subjects and 14.3 for the husbands of those who were or had been married; 61% of the husbands had jobs that were either professional/administrative or clerical/white collar, while the remainder were either in trades/skilled labor (22%), unskilled/semiskilled labor (3%), or unemployed/retired (14%); 21% of the subjects were employed; the average number of children for those ever married (98%) was 2. Comparison with U.S. Bureau of Census (1960) statistics showed these subjects to be slightly older, more likely to be married, better educated, having husbands who were not only better educated but of considerably higher socio-economic status, less often employed, and having more children than the average American female. Despite the rather biased nature of this self-selected sample, the test scores obtained by them had essentially the same means and variabilities as those designated as normative by the test publishers.

### Predictors

A large number of variables were included as possible predictors of accident behaviors. They were selected on the basis of a review of the literature in which all

seemed to have potential utility for prediction of accidents.

Measures of vision were made by means of the American Optical sight screener and included (a) acuity of right, left, and both eyes at both near and far distances, (b) stereopsis, near and far, (c) vertical phoria, near and far, and (d) lateral phoria, near and far.

Manual speed and dexterity were measured by the Employee Aptitude Survey Test No. 9 (EAS-9), a timed test of ability to place dots in small circles as accurately and quickly as possible. Intelligence was measured by the Otis Self-Administering Test of Intelligence—Higher Form, using a 20-minute time limit. Accuracy scores were derived from EAS-9 and the Otis by dividing the number right by the number attempted (right plus wrong).

Perceptual speed or attention to detail was measured by the Picture Completion subtest of the Wechsler Adult Intelligence Scale (WAIS).

Temperament trait measures were derived from the Guilford-Zimmerman Temperament Survey (GZTS), Guilford-Holley L Inventory, the DF Opinion Survey, Inventory of Factors GAMIN, and the Minnesota Multiphasic Personality Inventory (MMPI). They were as follows: (a) Impulsivity, (b) Emotional Stability, (c) Energy, (d) Aggressive Hostility, (e) Need for Independence, (f) Self-Confidence, (g) Dependency, (h) Intrapunitiveness, (i) Autism, (j) Cultural Conformity, (k) Meticulousness, (l) Hypochondriasis, and (m) Femininity.

Blood pressure was taken by an individual with nurse's training. Drinking habits, smoking habits, and use of drugs (depressants and stimulants) as well as health history and current health status variables were obtained by interview. Subjects were rated on weight, grooming, and nervousness. They were queried with respect to height, attitude toward housework, and demographic variables. Subsequent to kitchen performance, they were asked whether or not they were aware of having been observed. The speed with which they accomplished their tasks and the time of day (morning or afternoon) at which they participated were recorded by observers.

### *Observations*

All kitchen operations were observed by two individuals, each seated behind a one-way screen at opposite ends of the room. The observers remained silent throughout the operation, recording everything each subject did in the kitchen on a prepared checklist devised by staff in a preexperimental pilot study. Room was provided on the checklist for recording actions not listed. For each subject, interobserver reliability was computed and items on which the observers did not agree were disregarded in analyses. The mean interobserver reliability of 10 observers working in varying pairs over 6 months was .98. The observers also functioned as both interviewers and testers.

### *Procedures*

Each day of the 6-month data-collection period, two female subjects were scheduled for observation and testing, one in the morning and the other in the afternoon. The subject arrived at the laboratory, was greeted by an interviewer-observer, and sat at the kitchen table for her interview. She signed a waiver of legal responsibility and her blood pressure was taken. She was oriented to the kitchen by the interviewer, shown where all items to be used were located, given the list of tasks, and asked if there were any questions. Lists of items and their locations were posted for the subject's reference. When satisfied that the subject was ready to begin, the interviewer left the kitchen through the back, closed the door, and entered the observation room to begin his observation task. Meanwhile, the other observer was already seated at his window, unseen by the subject. The subject proceeded to carry out her assignments. When she was finished, an observer came into the room, immediately took her blood pressure, and then escorted her to the testing room where she was first given the simple reaction time test and the series of vision tests. She was then administered the Otis, EAS-9, and the Picture Completion test. At the end of on-site testing she was given a form containing items from the temperament scales and was told to take this form home, complete it without assistance, and mail it to the research office. Upon receipt of her form, she was to be mailed a check for \$10. No subject failed to return a completed test form.

### *Criteria*

Each subject received scores based on the number of (a) personal injury accidents, (b) property damage accidents, (c) total kitchen accidents (the sum of personal injury and property damage accidents), and (d) near accidents she was observed to have in the kitchen. Personal injury accidents consisted of "cuts," "jabs," "burns," "scalds," "bruises," and "falls." Bruises were considered analogous to "bumps" of a nature serious enough to bruise. Property damage accidents consisted of "breaks" (including "chips" or "cracks") or "burns" (including "melts" or "sets fire to") an object and "food spoilage" (as in dropping food on the floor or otherwise rendering it inedible). Some property damage was ascertained after the subject had left and after the damaged object was replaced by a new one. The near accidents category consisted of behaviors or events that, while they resulted in no injury or damage, fit the definition of unplanned events that logically could have resulted in an accident. Examples are seen in spilling of liquids or dropping of objects on the floor without removing them, failure to turn stove burners or the oven off at the end of kitchen performance, and loss of balance without falling.

In addition to kitchen accidents and near accidents, external criteria of automobile accidents and automobile violations were obtained from the California Department of Motor Vehicles for those 178 subjects who possessed driver's licenses. The project lasted sufficiently long to obtain records covering a 4-year period.

TABLE 1  
INTERCORRELATIONS OF ACCIDENT CRITERIA

Type of accident	1	2	3	4	5	6
1. Total kitchen accidents <sup>a</sup>	—	—	—	.31**	.16*	.19**
2. Personal injury accidents <sup>a</sup>			.24**	.30**	.16*	.13
3. Property damage accidents <sup>a</sup>				.18**	.08	.17*
4. Near accidents <sup>a</sup>					.20**	.09
5. Automobile accidents <sup>b</sup>						-.03
6. Automobile violations <sup>b</sup>						

Note. Relationships of personal injury and property damage accidents to total kitchen accidents are part-whole correlations.

<sup>a</sup>  $n = 226$ ; \*\*  $r = .17$  ( $p < .01$ ), \*  $r = .13$  ( $p < .05$ ).

<sup>b</sup>  $n = 178$ ; \*\*  $r = .19$  ( $p < .01$ ), \*  $r = .15$  ( $p < .05$ ).

### RESULTS

When kitchen accidents sustained by the 226 subjects were counted, the totals for each category were: (a) total kitchen accidents, 714; (b) personal injury accidents, 370; (c) property damage accidents, 344; and (d) near accidents, 648. It is important to note that in no case was an accident serious enough to interrupt kitchen performance for more than the length of time required by the subject to put on a bandaid. Automobile records for the subsample of 178 subjects showed that there were 36 automobile accidents and 108 automobile violations.

The unique property of the Poisson distribution is that its mean equals its variance. Therefore, the percentage of predictable variance in criteria is calculated by dividing the difference between the variance and the mean of the distributions by the variance ( $R^2 = (V - M)/M$ ) and multiplying by 100. The reliability of a given criterion is  $R$ . In the case of property damage accidents, the variance was less than the mean. For the remaining criteria, the percentages of predictable variance and reliabilities were as follows: (a) total kitchen accidents, 31%,  $R = .55$ ; (b) personal injury accidents, 24%,  $R = .49$ ; (c) near accidents, 14%,  $R = .37$ ; (d) automobile accidents, 20%,  $R = .45$ ; and (e) automobile violations, 16%,  $R = .40$ . These low reliabilities set a ceiling on the degree to which any criterion can correlate with any predictor. Using Cobb's (1940) approach to determination of the maximum possible correlation between each criterion and any "perfect" predictor provides an upper bound of .66 for total kitchen accidents, .61 for personal injury accidents, .40 for near

accidents, .50 for automobile accidents, and .45 for automobile violations.

Intercorrelations (Pearson  $r$  approximations) of criteria appear in Table 1. Where intercorrelations involve automobile accidents and automobile violations, they are based on data obtained from the 178 subjects with valid driver's licenses.

Significant relationships ( $p < .01$ ) were found between all kitchen-accident criteria as well as between near accidents and automobile accidents and between total kitchen accidents and automobile violations. The relationship between automobile accidents and personal injury accidents accounts for the correlation between automobile accidents and total kitchen accidents. This relationship is significant ( $p < .05$ ). It is interesting to note that (a) automobile accidents had no relationship to automobile violations in this sample, and (b) despite the fact that the property damage accident distribution fits the theoretical "chance" model, it correlates significantly with personal injury accidents, near accidents, and automobile violations. It should also be noted that some of these correlations may have failed to reach their potential values because of the characteristics of subjects who possessed driver's licenses. It was found that possession of a driver's license was negatively correlated with personal injury accidents ( $r = -.24$ ,  $p < .01$ ), total kitchen accidents ( $r = -.24$ ,  $p < .01$ ), and property damage accidents ( $r = -.13$ ,  $p < .05$ ). Thus, there is a restriction of range of kitchen accidents among those who have licenses that also restricts the correlations between automobile criteria and kitchen criteria.

Since there were a number of uncontrolled

TABLE 2

## SIGNIFICANT CORRELATIONS BETWEEN KITCHEN BEHAVIORS AND TOTAL KITCHEN ACCIDENTS

Behavior category and behavior item	<i>r</i>
Preparation (9 items)	
Makes use of correct tools (17 items)	none*
Uses paper cups in muffin tins	-.22**
Uses only rubber spatula in mixing	-.22**
Uses cutting board	-.15*
Uses paring knife for carrot sticks	-.15*
Uses correct soap to wash dishes	-.16*
Unsanitary practices (9 items)	
Eats or uses dropped food	.22**
Coughs without covering mouth	.23**
Uses same material to wipe floors and counters	.18**
Puts away a dirty object	.22**
Unsafe practices (31 items)	
Cuts bacon in hot frying pan	.17**
Uses cutting board in slot	.24**
Lays hot iron flat down	.17**
Handles plugged-in appliance with wet hands	.17**
Grabs sharp knife by blade	.17**
Climbs on object other than stepstool	.21**
Carries four eggs in hands	.20**
Pours bleach directly on clothes	.20**
Presses bacon with fingers in hot frying pan	.20**
Lets iron cord dangle on floor	.14*
Holds object in hand while climbing	.13*
Unplugs toaster by pulling on cord	-.19**
Wipes knife blade with fingers	-.15*
Safe practices (13 items)	
Turns iron off before unplugging	-.23**
Checks to see if beaters are firmly set before turning mixer on	-.16*
Unplugs iron immediately when through	.14*
Fails to follow directions (8 items)	none*

Note. *n* = 226 for total kitchen accidents.

\**p* < .05.

\*\**p* < .01.

variables that might have affected the outcome of the experiment, their effects were tested. The findings were as follows:

1. Variations in test administrators did not affect test results.

2. Fatigue effects were not evident in the comparison between morning and afternoon subjects. The only difference was that afternoon subjects had higher blood pressure both before and after the experiment than did morning subjects, but their performance was not affected.

3. The subjects who were aware that they were being observed had fewer accidents but

were also different from unaware subjects with respect to a number of accident-related variables and reported that, although aware of observers, they were so busy performing their tasks that they were rarely conscious of them.

One of the major objectives of the study was to explore the possibility of redefining accident criteria by including accident-related behaviors. Six categories of behavior were developed on the basis of observations. They were as follows:

1. *Preparation*—nine behaviors involving the subject recognizing something "wrong" (a condition contrived by the experimenters) at the start of her kitchen performance and remedying it (e.g., picking up a ball from the floor) or taking precautions to protect herself (e.g., putting on the apron).

2. *Makes use of correct tools*—consisting of 17 behaviors relating to selection of the correct alternative among objects provided for her use (e.g., using the cutting board to slice bread).

3. *Unsanitary practices*—consisting of 27 behaviors that were unhygienic (e.g., putting food that had dropped onto the floor into the lunch).

4. *Unsafe practices*—consisting of 31 behaviors that might lead to an accident (e.g., inserting a metal utensil into the toaster).

5. *Safe practices*—consisting of 13 behaviors that might avoid accidents (e.g., checking the setting on the iron before plugging it in).

6. *Fails to follow directions*—eight behaviors including any omission of an assigned task or deviation from assignment (e.g., not measuring ingredients according to recipe).

Each of the behaviors in these categories was correlated with all kitchen accident criteria. Table 2 shows the *r*s significant beyond the .05 level within each category and the total kitchen accidents criterion. In some cases, behaviors were related to other criteria but not to total kitchen accidents. Since total kitchen accidents is the most predictable criterion and since so many relationships were identified as to render their exposition beyond the scope of this report, it is used exclusively here.

Table 2 shows that none of nine possible relationships classified as preparation behaviors can be used for the prediction of total

kitchen accidents. When it comes to making use of correct tools, 5 out of 17 items are significantly correlated in the expected negative direction. Four out of nine unsanitary practice behaviors are significantly related in the expected positive direction. Unsafe practices provided two surprises in that 2 of the 31 possible relationships were in the negative direction (i.e., "unplugs toaster by pulling on cord" and "wipes knife blade with fingers"). However, 11 correlations are significant in the expected direction. Safe practices provided two correlations in the expected direction and one in the opposite direction. Following directions bore no significant relations to total kitchen accidents. The best behavioral predictors are in the categories labeled unsafe practices, unsanitary practices, and makes use of correct tools. Thus, while the results are not uniformly convincing, when one considers that none of these behaviors constituted any part of a criterion measure, it seems evident that there is a greater than chance relationship between some types of behaviors and accidents.

The last major objective of the study was to relate predictors to criteria. Only relationships between predictor variables and the total kitchen accident and automobile accident criteria are reported, the former because it is the major kitchen accident criterion and the latter because it is of greatest interest to accident researchers. Table 3 shows the significant ( $p < .05$ ) relationships.

Correlations were small, as expected, but all were in the direction predicted on the basis of previous research. None of the demographic variables was related to the automobile accident criterion, but women who were married (or had been), had children, and possessed driver's licenses had fewer total kitchen accidents. Personal handicaps and low blood pressure were positive predictors of automobile accidents. Use of tranquilizers or stimulants was positively related to total kitchen or automobile accidents, respectively, while the women who never drank alcoholic beverages seemed less likely to have kitchen accidents.

A number of visual measures were negatively related to total kitchen accidents, automobile accidents, or both. That is, good vision means less accident incidence. Intelligence bore a negative relationship to total kitchen acci-

TABLE 3

SIGNIFICANT CORRELATIONS BETWEEN PREDICTORS  
AND BOTH TOTAL KITCHEN ACCIDENTS (KA)  
AND AUTOMOBILE ACCIDENTS (AA)

Predictor	KA <sup>a</sup>	AA <sup>b</sup>
Demographic variable		
Marital status (ever married)	-.15*	
Number of children	-.21**	
Number of dependent children (under 18)	-.25**	
Driver's license	-.24**	
Health		
Handicap		.21**
Blood pressure—pretest		-.22**
Blood pressure—posttest		-.15*
Drug usage		
Takes tranquilizers	.14*	
Takes stimulants		.15*
Drinks alcoholic beverages	-.13*	
Vision: Visual acuity		
Binocular—far		-.17*
Binocular—near		-.20**
Right eye—far		-.15*
Right eye—near		-.19**
Left eye—near		-.15*
Both eyes—far		-.15*
Both eyes—near		-.21**
Stereopsis—near		-.17**
Cognition		
Otis number right		-.19**
Otis number wrong		.15*
Otis accuracy score		-.20**
Manual speed and dexterity		
EAS-9 number wrong	.15*	.20**
EAS-9 accuracy score	-.17**	-.16*
Attention to detail/perceptual speed		
Picture completion		-.19**
Temperament		
Need for freedom		.16*
Emotional stability		-.15*
Hypochondriasis		.17**
Self-reliance		-.15*
Nervousness (rating)		.15*
Kitchen performance		
Completes tasks		-.18*

<sup>a</sup>  $n = 226$   
<sup>b</sup>  $n = 178$ .

\*  $p < .05$ .

\*\*  $p < .01$ .

dents, and there was some confirmation of the hypothesis that errors are related to accidents, in that the accuracy scores on both the Otis and EAS-9 were negatively related to total kitchen accidents, and the EAS-9 accuracy

score was also negatively related to automobile accidents. Attention to detail and perceptual speed were also negatively related to total kitchen accidents.

Temperament measures or ratings did not fare well as predictors, nor were they consistent in their prediction of total kitchen or automobile accidents. The finding that completion of kitchen tasks was negatively related to automobile accidents and not to total kitchen accidents would seem surprising unless one considers that the more kitchen tasks performed, the greater the possible exposure to kitchen accidents.

Because age has been considered an important variable in accident research, its effect was evaluated by computing correlations ( $r_s$ ) with all accident criteria. None was significant. Eta coefficients were computed on the chance that the regression might be curvilinear but none were significant. In general, others have found consistent accident-age relationships for men but not for women. The results here would tend to support such previous findings.

### DISCUSSION

The results reported here do not include all of the analyses performed in the course of conducting the project from which they were derived. They do, however, reflect the major findings.

The most important aspect of the study was the structuring of a controlled environment within which it was possible to observe "accidents" (defined as unplanned harmful events) on the basis of human behaviors. The effect of a standardized environment is to provide control over exposure to hazard, thus rendering accident-causation attributable to human variability. The kitchen environment provided a natural setting for the 226 female subjects. The incidence of observed accidents (mishaps) was sufficiently high for meaningful analyses. That there was predictable variability in other criteria, including automobile accidents and violations, was demonstrated by calculation of the proportions of predictable variance and reliabilities for each criterion.

Significant intercorrelations among kitchen and automobile criteria served to confirm the hypothesis that accident incidence in one en-

vironment bears a positive relationship to accident incidence in other settings, a finding that some experts have contended supports the "accident-proneness" concept. Since the interrelationships between kitchen accidents and near accidents and automobile accidents and violations were restricted because of the tendency of women with driver's licenses to have significantly fewer kitchen accidents, the results are more impressive than the low correlations would suggest.

One of the major objectives of the study, identifying accident-related behaviors, was met with moderate success, in that it was possible to predict some criteria on the basis of some behaviors classified as (a) makes use of correct tools, (b) unsanitary practices, and (c) unsafe practices. Some behaviors seemed to relate in a manner contrary to that predicted, but in general, correlations were in the expected direction.

The results of correlating the predictors with kitchen accident and automobile accident criteria have been summarized. All correlations were low due to low reliabilities of criteria and, in the case of automobile accidents, restriction of range in other variables. However, they were of approximately the same magnitude as those found in most accident research and, in the kitchen laboratory, undistorted by nonbehavioral events.

The failure of some variables to demonstrate utility in this study is by no means critical, since the purpose of the effort was to broaden the definition of the term "accident" to include related human behaviors in the hope that psychologists concerned with the role of human error in accident causation will increasingly turn their attention to the patterns of human behavior which, if perpetuated, increase the probability of injurious or damaging consequences.

### REFERENCES

- ARBOUS, A. G., & KERRICH, J. E. Accident statistics and the concept of accident-proneness. *Biometrics*, 1951, 7, 340-432.
- BRODY, L. *Human factors research in occupational accident prevention: Its status and needs*. New York: New York University, American Society of Safety Engineers and Center for Safety Education, 1962.
- COBB, P. W. The limit of usefulness of accident rate as a measure of accident proneness. *Journal of Applied Psychology*, 1940, 24, 154-159.

- ENO FOUNDATION FOR HIGHWAY TRAFFIC CONTROL  
*Personal characteristics of traffic-accident repeaters.* New York: New York University, Center for Safety Education, 1948.
- FARMER, E., & CHAMBERS, E. G. *A psychological study of individual differences in accident rates.* (Industrial Health Research Board Rep. No. 38) London: Her Majesty's Statistical Office, 1926.
- FORBES, T. W. Analysis of "near-accident" reports. *Highway Research Board Bulletin*, 1957, 152, 23-37.
- GUILFORD, J. S. *An experimental study of home accident behavior.* (Final Report) Los Angeles: American Institutes for Research, December 1965.
- KEEHN, J. D. Factor analysis of reported minor mishaps. *Journal of Applied Psychology*, 1959, 43, 311-314.
- KRAFT, M. A., & FORBES, T. W. Evaluating the influences of personal characteristics on the traffic accident experience of transit operators. *Proceedings of the Highway Research Board*, 1944, 24, 278-291.
- KUNKLE, E. C. The psychological background of "pilot error" in aircraft accidents. *Journal of Aviation Medicine*, 1946, 17, 533-567.
- MCGUIRE, F. L. *A study of methodological and psychosocial variables in accident research.* (USPHS Final Report, Grant No. U100043) Irvine: University of California, Bureau of Community Environmental Management, March 1971.
- MINTZ, A., & BLUM, M. L. A re-examination of the accident proneness concept. *Journal of Applied Psychology*, 1949, 33, 195-211.
- NEWBOLD, E. M. *A contribution to the study of the human factor in the causation of accidents.* (Industrial Fatigue Research Board Rep. No. 34) London: Her Majesty's Statistical Office, 1926.
- RUCH, F. L., & WILSON, C. L. A new system for selecting safe drivers. *Commercial Carrier Journal*, 1948, 75, 66-68.
- THORNDIKE, R. L. *The human factor in accidents with special reference to aircraft accidents.* (Project No. 21-30-001, Rep. No. 1) Randolph Field, Tex.: USAF School of Aviation Medicine, February 1951.
- VASILAS, J. N., FITZPATRICK, R., DUBois, P. H., & YOUTZ, R. P. *Research on near accidents.* Pittsburgh: American Institutes for Research, September 1952.
- WHITLOCK, G. H., CLOUSE, R. J., & SPENCER, W. F. Predicting accident proneness. *Personnel Psychology*, 1963, 16, 35-44.
- WONG, W. A., & HOUBBS, G. E. Personal factors in industrial accidents—a study of accident proneness in an industrial group. *Industrial Medicine*, 1949, 18, 291-294.

(Received February 3, 1972)

## THE RECOGNITION OF ROAD PAVEMENT MESSAGES<sup>1</sup>

WENDY A. MACDONALD

*Australian Road Research Board  
Melbourne, Australia*

ERROL R. HOFFMANN<sup>2</sup>

*Department of Mechanical Engineering  
University of Melbourne*

The relationship between recognition threshold and degree of elongation of letters used in road pavement messages was investigated. Experiments were conducted in the laboratory and in a field situation. It was found that in both situations the normally proportioned letters were recognized at smaller visual angles than the more elongated letters; increases in letter elongation did not produce increases in recognition distance directly proportional to the increases in the vertical visual angle subtended. Mathematical models based on the relationship between perceived and real distance largely describe the observed effect, and a formula is given by which traffic engineers can calculate the necessary degree of letter elongation for a desired threshold recognition distance.

Highway engineers often use word messages painted on the road surface, such as No Right Turn, Left Turn Lane, or Ped X-ing Ahead. The letters in these words are elongated according to instructions in the manuals of standard practice of the road authorities. The Standards Association of Australia Road Signs Code (Standards Association of Australia, 1960a) states that "The letters should be greatly elongated in the direction of traffic movement because of the low angle at which they are viewed by approaching drivers. The 'height' of roadway lettering will depend on prevailing traffic speeds, but, in any case, they shall not be less than 6 feet long [p. 121]."

No particular basis for this policy is stated, but there appears to be an underlying assumption that the more elongated are the letters, the greater is the distance at which the message can be read. No limiting condition is suggested. And yet there are practical reasons why it is important that degree of letter elongation should be kept to a minimum: Economically, it is undesirable to use time and paint applying unnecessarily large messages; and from a safety viewpoint, there is evidence that at least some of the commonly used paints have lower coefficients of friction than unpainted road

surface, with a consequent rise in the risk of skidding.

Therefore, the general purpose of this study was to determine the gains in threshold recognition distance that might be expected with increased letter elongation. The approach was simply to establish the form of the relationship between recognition threshold and vertical visual angle subtended for letters of varying elongations.

If recognition distance is completely determined by visual angle subtended, then only the law of the visual angle is operative. In this case, assuming that recognition takes place at some fixed value of visual angle, it would be necessary to increase letter elongation proportionally to the square of the distance at which recognition was required. From this it follows that, in terms of the geometry of the situation, there is a practical limit to the benefits to be gained by elongating letters.

In addition to this known geometrical relationship, the presence of additional, complicating factors was suspected, whose effects would be to modify the operation of the law of the visual angle so that the advantage of greatly elongated letters would be decreased even further. For instance, it was thought that the phenomenon of perceptual constancy (Day, 1969; Epstein, Park, & Casey, 1961; Holway & Boring, 1941; Thouless, 1931a, 1931b) might differentially affect the recognition thresholds of letters of different elongations, leading to a relative disadvantage for the more elongated letters.

<sup>1</sup> Financial and other support was provided by the Australian Road Research Board.

<sup>2</sup> The authors wish to acknowledge the helpful suggestions of C. Cameron during the course of this work.

Requests for reprints should be sent to E. R. Hoffmann, Department of Mechanical Engineering, University of Melbourne, Parkville, Victoria 3052, Australia.

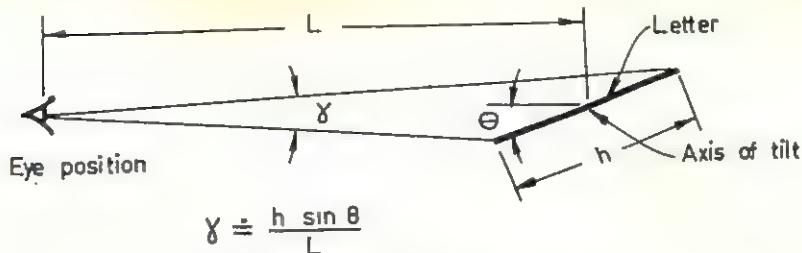


FIG. 1. Geometry of laboratory experiment.

The hypothesis under test was a simple one: that recognition of letters of varying elongations occurs at a constant subtended vertical visual angle. The approach was empirical, with no attempt made to clarify causal mechanisms. Nevertheless, the results are discussed in terms of two mathematical models based on the known effects of perceptual constancy, which predict the effect on recognition threshold of perceived as opposed to real letter height.

#### METHOD

Two experiments were carried out: (a) in the laboratory, where viewing distance was held constant and subtended visual angle was varied by tilting the letter about an axis in the fronto-parallel plane to the observer (see Figure 1) and (b) in the field, where subtended visual angle was varied by changing the viewing distance (see Figure 2).

The experimental methods differed primarily in that laboratory viewing distance was held constant, whereas in the field, as in "real life," it varied. The laboratory method was adopted as being the most suitable for a brief initial check on the experimental hypothesis, and the results of this check clearly established the need for the full-scale field experiment.

The equations for the subtended visual angles under these two sets of experimental conditions are given in Figures 1 and 2. Both equations assume that degree of elongation (referred to throughout the article as letter height,  $h$ ) is very small in relation to viewing distance,  $L$ .

#### Laboratory Experiment

*Subjects.* These were 12 male members of staff of the Department of Mechanical Engineering, in the age range 20 to 40 years, with normal eyesight.

*Procedure.* White letters on a grey background were displayed one at a time on a large board. Letter luminance was  $3.4 \text{ cd/m}^2$  and background luminance was  $1.0 \text{ cd/m}^2$ . The letters were pivoted about an axis on the line of sight of the seated subject (see Figure 1). A chin rest was used to maintain the subject's head in the correct position. Viewing distance,  $L$ , was 7 feet. The letters were from the standard set recommended for use in pavement messages (Standards Association of Australia, 1960b). Three alphabets were used, having heights of 2 inches, 3.7 inches, and 5.4 inches. (The term alphabet is used throughout the article to refer to a group of letters having a common height.) Mean letter width was constant for all three alphabets at 0.8 inches. To avoid unduly long experimental sessions, only 12 letters per alphabet were used. These were A, B, C, F, H, J, L, N, S, T, W, and Y. Order of presentation was balanced over subjects. Each subject was given 72 practice presentations of the letters before commencing the experiment.

Each letter was presented at one angle only per presentation. In choosing the initial angles of presentation for the session, half of the letters were presented at angles estimated to be above threshold for those particular letters, and half were presented at angles below their estimated thresholds. Presentations were repeated at different angles of tilt until recognition occurred (for those letters which were initially below threshold) or ceased to occur (for those letters which were initially above threshold). The performance for each alphabet was then taken as the mean angle, over

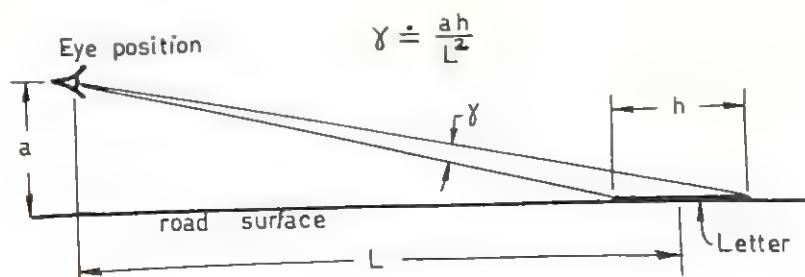


FIG. 2. Geometry of field experiment.

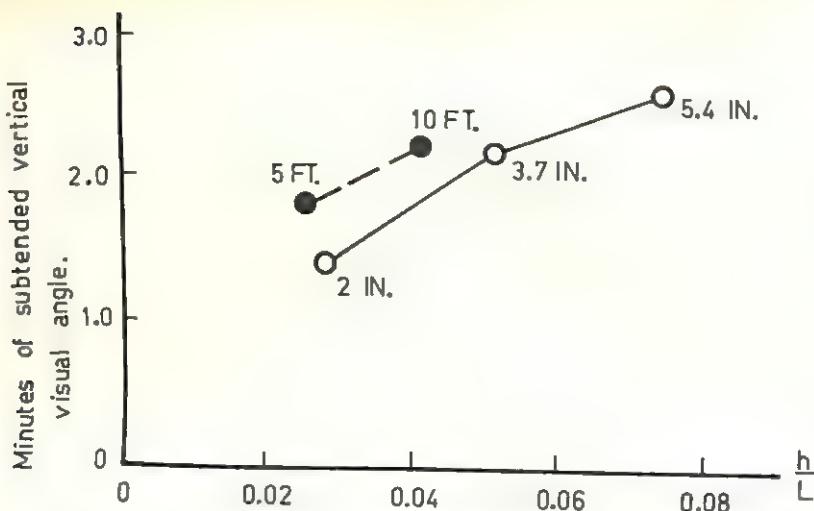


FIG. 3. Variation of subtended vertical visual angle at recognition with height-to-distance ratio of the letter.

all letters in that alphabet, at which this change in response occurred.

#### Field Experiment

**Subjects.** These were 12 undergraduate students, 5 males and 7 females, of normal eyesight.

**Procedure.** Two 16-letter alphabets were used, one 5 feet high and one 10 feet. Each alphabet had a mean letter width of 2 feet 3 inches. The letters were painted in white on grey. Those used were A, C, D, E, F, G, H, I, L, N, O, P, R, S, T, and U, taken from a standard set of letters similar in shape to those used in the laboratory experiment (National Association of Australian State Road Authorities, 1970). To facilitate the construction of these full-scale letters the geometrically simplest were used, resulting in a different subset from that used in the laboratory. However, neither subset would be expected to introduce any significant bias related to shape of individual letters (Cornog & Rose, 1967, pp. 162, 201, 251, 284). Letter luminance was around 19,000 cd/m<sup>2</sup> and background luminance was around 1,700 cd/m<sup>2</sup>.

Letters were laid down, one at a time, on an airport runway. The subject sat in the driver's seat of a Ford Fairlane sedan. He observed the letters through a slit at a standard height of 39 inches above ground level. The side of the runway was marked in increments of visual angle subtended by the letters. Starting from the most distant mark (0.25 minutes of visual angle) the experimenter drove the car towards the letter, stopping at each mark, where the viewing slit was uncovered and the subject attempted to recognize the letter. If he responded wrongly he was told so and driven to the next mark. On correct recognition of the letter the car was returned to the starting mark for the next letter. Order of presentation was balanced over subjects.

#### RESULTS

Data averaged over all subjects and letters are presented in Figure 3. Laboratory performance was recorded in terms of threshold angle of tilt, while field performance was recorded in terms of threshold distance, but both these measures were converted to threshold vertical visual angle subtended by the letters. There were also differences in scale between the experiments, but the ratio of letter height to viewing distance ( $h/L$ ) was the major geometrical ratio in both. By the principles of dimensional analysis the two sets of data are therefore directly comparable in the form shown in Figure 3.

It can be seen that, in the laboratory, subtended vertical visual angle at recognition increased with letter elongation (analysis of variance,  $F = 4.14$ ,  $df = 2/33$ ,  $p < .05$ ). The field experiment showed a similar effect (median test,  $\chi^2 = 10.01$ ,  $df = 1$ ,  $p < .01$ ).

#### DISCUSSION

In neither situation did recognition occur at a constant subtended visual angle. The generally higher threshold values in the field experiment compared with the laboratory are probably related to the fact that in the field the subject always moved from below threshold up to threshold, whereas in the laboratory half of the approaches were from above threshold.

But overall it is evident that the two sets of data are in reasonable agreement.

The visual angles for the field study correspond to recognition distances of 188 feet for the 5-foot letters and 241 feet for the 10-foot letters. That is, there was a gain of 53 feet in the recognition distance for 10-foot over 5-foot letters. If the 10-foot alphabet had been recognized at the same visual angle as the 5-foot alphabet, this gain would have been 76 feet. It should be remembered that these threshold distances are for single letters, not words.

#### *Mathematical Models*

It has been firmly established (e.g., Day, 1969; Thouless, 1931a) that under normal viewing conditions an observer's judgments of object size, shape, etc. vary much less than the corresponding aspects of subtended visual image. "This *relative* stability of perceptual judgments of object characteristics with variation in their sensory representations is called perceptual constancy [Day, 1969, p. 58]." In the introduction it was suggested that perceptual constancy might affect recognition thresholds; that is, people might find it easier to recognize a letter which is perceived as nearer, or larger, or perhaps more normal in shape than another, even although both subtend the same visual image.

In the present situation, where subtended vertical visual angle is very small, height effects are likely to be more important than those of width. The phenomenon of distance constancy means that equal increments of distance appear smaller as the absolute distance from the observer is increased. Considering letter height as an increment of distance, the extent of any constancy effect in relation to letter height may be predicted by either of two simple mathematical models derived from the experimentally established relationship between perceived and real distance (Gilinsky, 1951; Stevens, 1957). These models describe the relationship between letter height and recognition threshold, where viewing distance is much greater than letter height.

*Stevens' power law model.* According to Stevens' (1957) law, the relationship between perceived (subscript  $p$ ) distance and real distance is given by

$$L_p = KL^n \quad [1]$$

where  $K$  is a constant and the index  $n$  is shown by Teghtsoonian (1971) to be dependent on the stimulus range ratio in the particular experiment.

Regarding letter height as a component of perceived distance and referring to Figure 2, the perceived height of the letter ( $h_p$ ) is given by

$$h_p = K(L + h/2)^n - K(L - h/2)^n \quad [2]$$

Expanding by use of the binomial theorem and neglecting small terms gives

$$h_p = KnL^{n-1}h \quad [3]$$

Substituting for  $h_p$  in the relationship  $\gamma_p = ah_p/L^2$  (see Figure 2) gives the perceived visual angle

$$\gamma_p = KL^{n-3}ah \quad [4]$$

If the perceived visual angles of letters of real heights  $h_1$  and  $h_2$  are equal, it is assumed that they would have the same probability of recognition. With this assumption, the corresponding recognition distance ratio is given by

$$\frac{L_1}{L_2} = \left( \frac{h_2}{h_1} \right)^{1/n-3} \quad [5]$$

When  $h_2 = 2h_1$  (as in the field experiment), and with  $n = 0.67$  (a value quoted by Stevens), this model predicts that  $L_2 = 1.34L_1$ , in comparison with the experimentally obtained relationship of  $L_2 = 1.28L_1$ . If it is assumed that recognition occurs when real visual angles are equal (law of the visual angle) the relationship is  $L_2 = 1.41L_1$ .

*Gilinsky's retinal model.* An alternative model may be based on the equation given by Gilinsky (1951) relating perceived and real distance; that is,

$$L_p = \frac{AL}{A + L} \quad [6]$$

where  $A$  is an experimentally determined constant which is dependent on the viewing conditions. A procedure similar to that used above with Stevens' (1957) formulation gives, for equal probabilities of recognition,

$$L_p^2 h_1 (A + L_2)^2 = L_1^2 h_2 (A + L_1)^2 \quad [7]$$

When  $A = 130$  feet (a value found by Gilinsky) and for the same conditions as those in the field experiment, this equation predicts

a value of  $L_1 = 194$  feet when  $L_2 = 241$  feet, which is only slightly greater than the experimentally obtained value of  $L_1 = 188$  feet. The law of the visual angle predicts a distance of  $L_1 = 171$  feet.

*Model accuracy.* Both models predict ratios of recognition distances for the two alphabets, the values of which are close to those found in the experiment. In other words, the models predict with fair accuracy the extent of the observed deviation from the law of the visual angle.

This level of accuracy seems a little surprising in view of the fact that both models make use of constants whose values depend on the effective stimulus range for the particular experiment (Teghtsoonian, 1971)—values which in this case are unknown. Therefore the values of constants used in the above calculations were ones found by Stevens (1957) and Gilinsky (1951), which are not necessarily applicable to the present situation. Also, letter shape was an uncontrolled factor in this study; mean horizontal visual angle was the only "shape" parameter held constant over alphabets. Two letters may subtend the same horizontal visual angle at mid-height but the more elongated letter will have a greater difference between the horizontal visual angles subtended at its near and far edges than the other letter.

The accuracy achieved in spite of these potential sources of error strongly suggests that size/distance constancy was a major factor affecting letter recognition.

### Design Equation

The main purpose of this study was to determine the benefits of increased letter elongation in terms of increases in threshold recognition distance. The present design manual (SAA, 1960a) simply says that "letters should be greatly elongated [p. 121]." This study has provided a basis for a quantitative formula to replace the above instruction. This formula could be used by traffic engineers to calculate necessary letter elongations to achieve desired recognition distances.

Such a formula may be derived as follows: assuming the relationship expressed in Equation 4 above, and calculating the constants from the experimental data relating to the 10-foot

letters, then

$$L = 54.2(ah)^{0.43} \quad [8]$$

The validity of this equation for design purposes was checked in an additional experiment on the recognition of real pavement messages (Macdonald & Hoffmann, in press). A least-squares fit to this "real world" data gave the equation

$$L = 45.6(ah)^{0.44} \quad [9]$$

From the similarity of the indices in Equations 8 and 9 it appears that the general form of the equation is valid, with Equation 9 being the more suitable for practical use. It is evident from the different values of the constants that recognition threshold distances were greater for single letters than for real messages. These messages varied in height from 2 to 22 feet, so the equation is applicable to at least the range of letter heights likely to be used in practice. Other aspects of pavement messages, such as word order and spacing, are discussed by Macdonald & Hoffmann (1972).

### CONCLUSIONS

These findings indicate:

1. Elongated letters were recognized at a greater distance than "shorter" letters of the same width, but the increase in recognition distance was significantly less than would be expected if vertical visual angle subtended were the only determining factor.

2. Simple mathematical models based on the relationship between perceived and real letter heights predict the experimental results fairly well, suggesting that perceptual constancy might have affected recognition thresholds.

3. An equation has been found that appears suitable for calculating the necessary letter height to achieve a desired recognition distance.

### REFERENCES

- CORNOC, D. Y., & ROSE, F. C. *Legibility of alphanumeric characters and other symbols. II. A reference handbook*. United States Department of Commerce, National Bureau of Standards, Miscellaneous Publication 262 2, 1967.
- DAY, R. H. *Human perception*. Sydney, Australia: Wiley, 1969.

- EPSTEIN, W., PARK, J., & CASEY, A. The current status of the size-distance hypotheses. *Psychological Bulletin*, 1961, 58, 491-514.
- GILINSKY, A. S. Perceived size and distance in visual space. *Psychological Review*, 1951, 58, 460-482.
- HOLWAY, A. H., & BORING, E. G. Determinants of apparent visual size with distance variant. *American Journal of Psychology*, 1941, 54, 21-37.
- MACDONALD, W. A., & HOFFMANN, E. R. Factors affecting the design of road pavement messages. *Proceedings of the Sixth Conference of the Australian Road Research Board*, in press.
- NATIONAL ASSOCIATION OF AUSTRALIAN STATE ROAD AUTHORITIES. *Guide to the publications and policies of NAASRA*. Sydney, Australia: Author, 1970.
- STANDARDS ASSOCIATION OF AUSTRALIA. *Australian standard rules for the design, location, erection, and use of road traffic signs and signals*, Australian Standard No. CE. 1-1960. Sydney, Australia: Author, 1960. (a)
- STANDARDS ASSOCIATION OF AUSTRALIA. *Standard alphabets for road signs*, Australian Standard No. E.37-1960, Series A. Sydney, Australia: Author, 1960. (b)
- STEVENS, S. S. On the psychophysical law. *Psychological Review*, 1957, 64, 153-181.
- TEGHSOONIAN, R. On the exponents in Stevens' law and the constant in Ekman's law. *Psychological Review*, 1971, 78, 71-80.
- THOULESS, R. H. Phenomenal regression to the real object: I. *British Journal of Psychology*, 1931, 21, 339-359. (a)
- THOULESS, R. H. Phenomenal regression to the real object: II. *British Journal of Psychology*, 1931, 22, 1-30. (b)

(Received November 23, 1971)

## THE DYNAMIC ROLE OF EYE-HEAD ANGULAR DISPLACEMENTS IN HUMAN VEHICULAR GUIDANCE<sup>1</sup>

HENRY S. R. KAO<sup>2</sup>

*Highway Safety Research Institute, University of Michigan*

This research reports three experiments on the effects of angular displacements of driver's vision, head, and combined eye-head separation on steering produced through closed-circuit television systems. Steering errors increased with increasing magnitudes of such displacements in actual driving. Results are discussed with respect to driver-vehicle visual requirements, visual-motor coordination in steering, and design of driver training devices.

The driving process may be viewed as a closed-loop, feedback-controlled, driver-vehicle-road tracking system with well-defined target, cursor, control, and driver interrelationships (Gordon, 1966; Kao & Nagamachi, 1969a; Kao & Smith, 1969). Crucial to this system is the role of an individual as a control system which generates a course of action through steering, and controls and corrects the consequent vehicle movements by means of sensory feedback (Hendricksson, Nilsson, & Anderson, 1965; Smith & Smith, 1966).

Within the context of driver-vehicle-road tracking, an automobile may be conceived to be an exoskeletal machine, with the human factors in steering control defined primarily by the operational relationships between the slave skeleton and the driver's movements in operating controls from a personal reference. The effectiveness of this driver-vehicle system is determined by the dynamic space, time, force, and feedback of eye, head, and postural movements as well as the articulated motions of the extremities. The basic considerations here are the feedback transformations between the driver's control actions and sensory information and the effects of these actions as communicated by the machine skeleton

(Mosher, 1965; Mosher & Knowles, 1961; Smith, 1966). Efficient driving performance depends upon refined motor control to guide the steel shell in tracking a well-defined road and in controlling the effects of vehicle action on the sensorimotor system (Kao, 1969; Kao & Smith, 1969).

The fundamental sensorimotor control system within this master-slave relationship in driver-vehicle-road tracking depends largely upon five levels of feedback referencing for optimal driving efficiency: eye position, head position, postural position and motion, vehicle positioning, and road geometry. The hypothetical zero reference for road tracking is the optimal alignment of vision, head, upright seated posture, and vehicle positioning relative to the road on a parallel longitudinal plane. For dynamic steering control, these levels of information are continuously displaced at varying degrees from the changing zero reference plane dependent upon the road geometry, driver motor refinement through practice, vehicle movements, positional changes, and other conditions. These dynamic sensory displacements constitute the driver's feedback information for subsequent error detection, correction, and further initiation of motor control (Kao, 1969).

One way to investigate the assumption of eye, head, and posture as the basic referencing systems for driving performance is to modify experimentally the sensorimotor control loop between the driver and the automobile exoskeleton in order to measure the effects of these relationships on steering control. Displacement of sensory feedback has been the

<sup>1</sup> This paper was originally presented at the ERS-IEEE International Symposium on Man-Machine Systems, Cambridge, England, September 8-12, 1969. The research reported was supported by the Highway Safety Research Institute of the University of Michigan and was completed there.

<sup>2</sup> Requests for reprints should be sent to Henry S. R. Kao, who is now at the Psychology Department, Glassboro State College, Glassboro, New Jersey 08028.

most commonly used modification in the study of human perceptual and motor behavior. Research on displacement of vision in driving is a natural extension of previous studies of the effects of angular and lateral displacement of vision in human performance and learning studies of spatial orientation (Kohler, 1955; Smith & Smith, 1962; Smith, Smith, Stanley, & Harley, 1956; Stratton, 1897; Wooster, 1923).

Kao and Smith (1969) have recently reported a study using a closed-circuit television system to investigate the effects of laterally displaced vision of the vehicle's forward view on the accuracy of vehicle guidance. When driving a car with a substitute television image of the left, center, and right sections of the hood obtained by laterally shifting the camera position on top of the car, the subject's road tracking was found to be most efficient when the view of the center of the hood was presented. The view to the left of the front hood was shown to be superior to that of the right portion in providing information necessary for effective vehicle control and guidance on the road.

This study extends the initial feedback research on displaced vision in steering by three further investigations on the effects of angular displacement of eye and head and angular separation of eye and head on driver steering performance. Based on the feedback referencing concepts, the following were assumed to degrade steering accuracy: (a) angular displacement of driver's vision from the longitudinal plane, (b) angular displacement of the head position from the longitudinal plane, and (c) combined eye-head angular separation. Each assumption was tested in one experiment. Results are discussed in terms of the theoretical framework, the driver's vision of vehicle features, and their implications in driver skill learning and training.

#### EXPERIMENT I: ANGULAR DISPLACEMENT OF VISION AND DRIVER STEERING PERFORMANCE

##### Method

**Subjects.** Twelve male and female students and staff members from the University of Michigan were used as subjects. All subjects had driver's licenses. Student subjects were paid for their participation.

**Test vehicle.** A 1967 Plymouth Fury 4-door sedan was used. It had automatic transmission and power steering, an overall length of 17 feet 7 inches and front bumper width of 6 feet 5 inches.

**Task course.** The task course was a slightly winding roadway resembling two complete sine waves and marked with red 14-inch reflective traffic cones at fixed intervals of 10 feet. The width of the course was 8 feet, with a total length of 255 feet. It was made continuous by connecting the cones with 4-inch white traffic tapes.

**Speed.** The speed was preset at a constant 15 miles per hour throughout the experiment.

**Displaced vision conditions.** To produce displaced vision, a Sony model CVM-51UWP 8.5-inch transistor television monitor and an Ampex model C-Mount CC-324 camera with 25-millimeter Vidicon lens were used. A Topaz model 310-B-12, 300-W inverter provided power from the engine battery. In the actual testing, the television monitor was located constantly in front of the steering wheel on the outside of the windshield facing the driver. The camera was placed on the roof of the car corresponding to the driver's head position (see Figure 1).

The camera was placed on the car roof directly over the driver's head with horizontal and vertical angles of 28° and 21°, respectively, on a specially designed wooden carrier with variable height and longitude. The angular displacements of vision were achieved by placing the camera straight forward along the longitudinal plane of the vehicle (zero displacement) and by 10° and 20° camera rotations off to the right. These camera positions gave monitored images of the left half, center, and right half of the hood with a longitudinal display occupying approximately two thirds of the monitor screen. The monitor was fixed in position throughout the experiment.

The windshield, rear window, side windows on the driver's side, and the front window on the side opposite the driver were completely covered with heavy blankets.

**Design and procedure.** A Latin-square design was used to assign the order of treatments to subjects for the six combinations of three camera positions. This was repeated once for each of the 12 subjects. There were 5 runs for each treatment, making a total of 15 runs per subject.

The subjects' task was to drive the test car through the course at a fixed speed of 15 miles per hour by viewing only the television display.

Before the experiment, each subject was told the nature of the study and task requirements. Preceding each camera condition, each subject was given four practice runs through a straight course of 50 feet marked with cones at the same intervals and width as the test course. Subjects were then instructed to drive the car through the task course from a starting line 300 feet away from the first cones. For the practice and experimental trials, the subjects were guided to the right track by verbal instructions given by an experimenter sitting in the rear seat looking out of the window. The subject was completely on his own

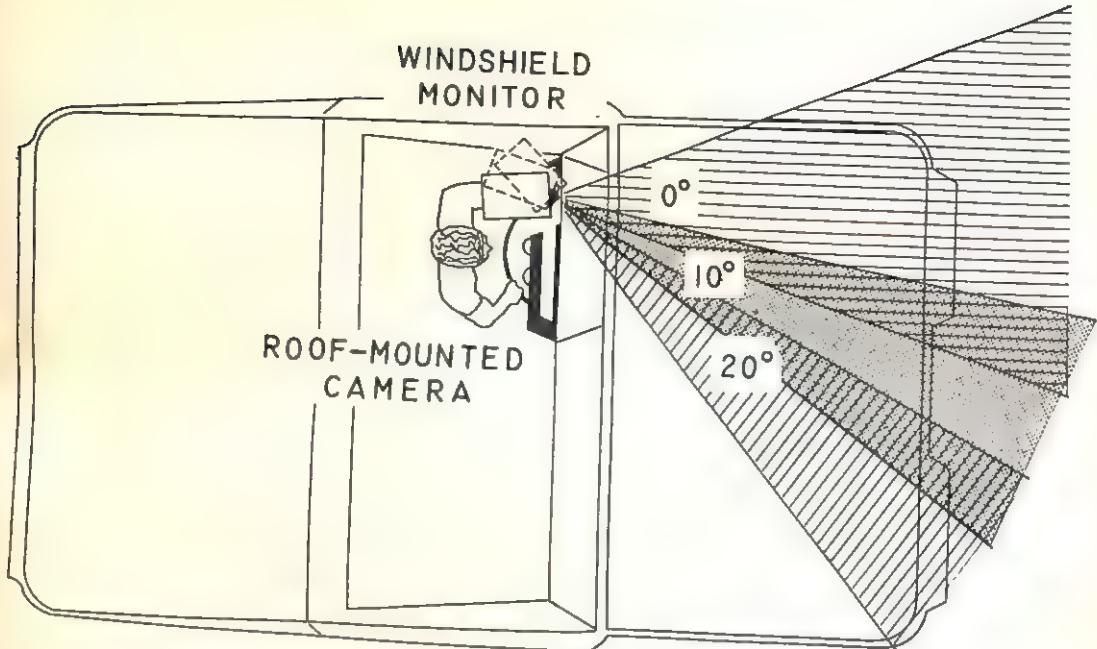


FIG. 1. Experimental television system and the arrangement of visual angular displacement conditions.

after he was directed to the right course at the fixed distance.

Use of seat belts was required of all subjects, and use of the foot brake during the task was forbidden. Tracking error was measured by the number of traffic cones touched or knocked down in each trial. The error score of each test trial was used as the basis for an analysis of variance. Due to the alternate order of treatments and changing of camera positions, each subject had rest periods of approximately 20 minutes between conditions. In driving with the substitute television images, the driver maintained the line of sight normally used in unaltered vision of the road and normal head and posture positions. The images obtained gave a fairly clear indication of the test course some 75 feet ahead of the car.

### Results

Analysis of variance and Duncan multiple-range tests were carried out to assess the statistical significance of the displaced vision conditions.

The tracking error was least for the straightforward camera position with a television image of the left half of the vehicle hood and increased as a function of increased angular displacement of vision in terms of camera rotation from the zero longitudinal plane. Performance was poorest when the right half of the vehicle hood was displayed ( $20^\circ$  camera rotation). Mean task errors

were 6.30, 7.30, and 12.30 for visual displacements of  $0^\circ$ ,  $10^\circ$ , and  $20^\circ$ , respectively.

Results of a two-way analysis of variance show a significant difference between the three visual displacement conditions ( $F = 24.08$ ,  $df = 2/22$ ,  $p < .001$ ).

Results of the Duncan range test show that the visual displacement of  $20^\circ$  rotation with a monitor image at the right hood was significantly different from the center and left images ( $p < .01$ ). The center image was not significantly different from the left, under which the best road tracking performance was obtained.

Learning across trials under each condition is shown in Figure 2. Mean task errors of six different subjects in a separate pilot study under normal visual conditions and like procedure are also presented in this figure and Figures 4 and 5 later for comparison purposes. Except for the right-image condition where the first trial is significantly different from all the rest at the .01 level, steering performance under the displaced vision conditions resulted in no significant improvement across trials, as separate Duncan multiple-range tests have shown. In other words, no learning was statistically demonstrated in the driving task under angularly displaced conditions of vision.

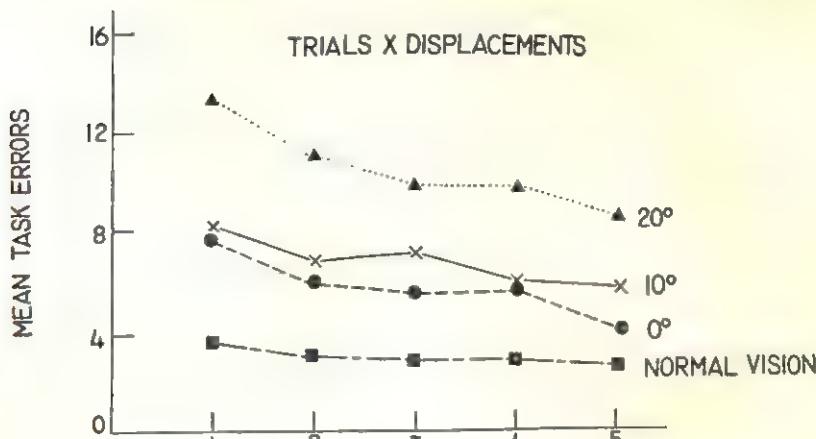


FIG. 2. Driving performance across trials under normal and each condition of visual displacement.

## EXPERIMENT II: ANGULAR HEAD DISPLACEMENTS AND DRIVERS STEERING PERFORMANCE

### *Method*

Method, design, and procedure used in this experiment were identical to those in Experiment I, except for the experimental conditions of head displacement.

For the design of head angular displacements, three Sony model CVM-51UWP 8.5-inch transistor television monitors were placed immediately outside of the windshield in positions corresponding to the center line of the hood and the midpoints between this line and the two lateral edges of the hood. They represent points facing approximately the driver's and right passenger's line of sight and the center of the vehicle. With the left monitor as the zero reference line, the monitor positions represent medial rota-

tions of  $0^\circ$ ,  $32^\circ$ , and  $51^\circ$  from the driver's frontal head line. For each angular displacement, the screen of the monitor was so positioned that when the driver's head was turned to the respective directions, perpendicular viewing was obtained.

The television camera was placed on top of the car so that it corresponded to the driver's head position and remained constant with a straight coverage. With this coverage, the monitor displayed the road course about 75 feet ahead of the vehicle and a constant view of the left half of the vehicle hood. Under this arrangement, the subject had to turn his head toward the monitor direction selected for each trial (see Figure 3).

Five trials for each of the three head displacement conditions were run for each of the 12 subjects, preceded by four practice trials on a straight course, both under the initial guidance of an experimenter.

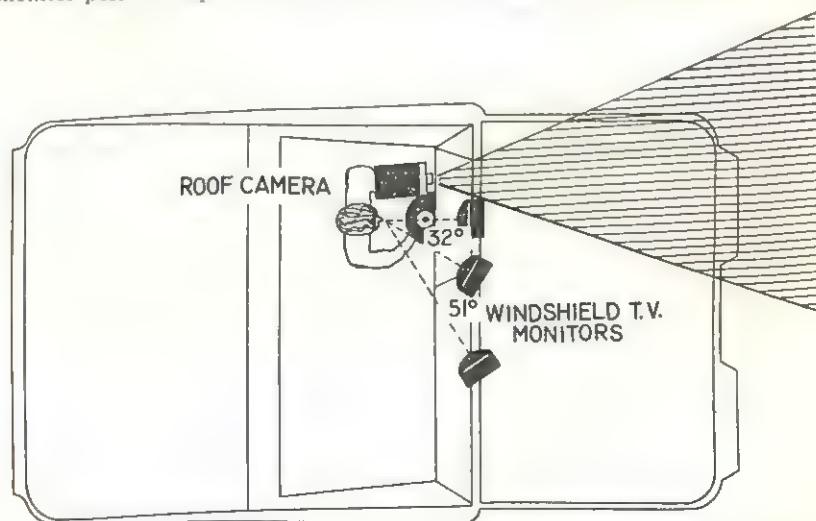


FIG. 3. Experimental television system and the arrangement of angular head displacement conditions.

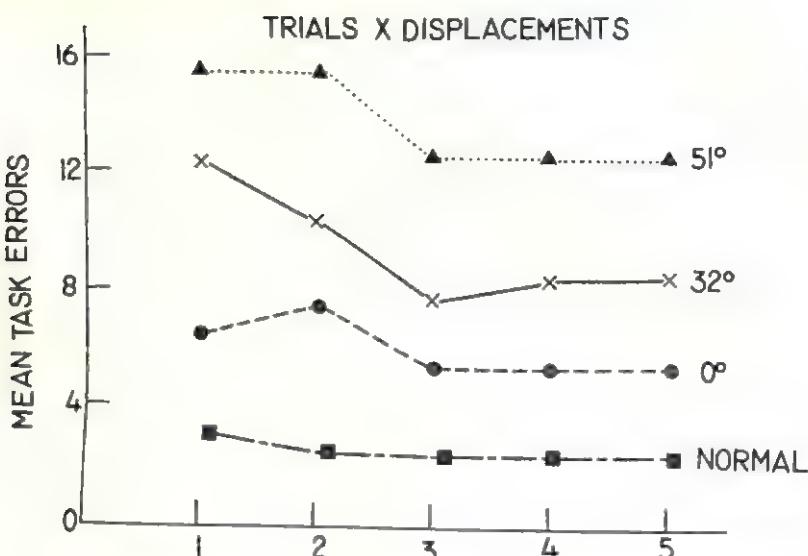


FIG. 4. Driving performance across trials under normal and each condition of head displacement.

## Results

The mean performance efficiency in this driving task under three levels of head displacement for all subjects is shown to be 6.01, 9.30, and 13.62 for head displacements of 0°, 32°, and 51°. The mean tracking errors were found to increase as a direct function of increase in the angular displacement of the head off the normal longitudinal plane (0° displacement). Most efficient performance was recorded in the 0° head displacement condition. Results of analysis of variance show a significant difference in the three displacement conditions of head position ( $F = 16.97$ ,  $df = 2/22$ ,  $p < .001$ ). This difference between the three specific conditions was further substantiated by a Duncan range test at the .01 level. An overall difference between trial conditions was found in the analysis ( $F = 8.27$ ,  $df = 4/44$ ,  $p < .001$ ).

Performance across trials under each of the experimental conditions is illustrated in Figure 4. The results of an analysis of variance did not show statistical difference for the overall Displacement  $\times$  Trial interaction. No significant difference was found between the 0° head displacement condition and the 0° eye displacement condition in Experiment I.

Driving performance under a normal vision condition with identical experimental procedures obtained from previous data is again plotted at the bottom of Figure 4. As men-

tioned earlier, no motor learning or improvement was observed across five experimental trials.

## EXPERIMENT III: COMBINED ANGULAR EYE-HEAD SEPARATION AND DRIVER STEERING PERFORMANCE

This experiment was designed to further test the effects of combined eye and head separation in human vehicle control and guidance. This was achieved by introducing a constant visual angular displacement of 20° to the right of the longitudinal reference plane in the three head displacement conditions reported in Experiment II. From the feedback viewpoints of driving control, it was assumed that (a) angular head displacement together with constant eye displacement forming a combined eye-head separation would be best for driving control when such a separation is minimized and (b) the decrement in performance would be a function of increased angular separation between the eye and head displacements. This was predicted in spite of the fact that this condition represented a fundamental shift of the eye-head alignment to the right in the normal driving situation.

## Method

The design and experimental procedure of this experiment were identical to those of Experiment II, except that the camera was rotated a constant 20°

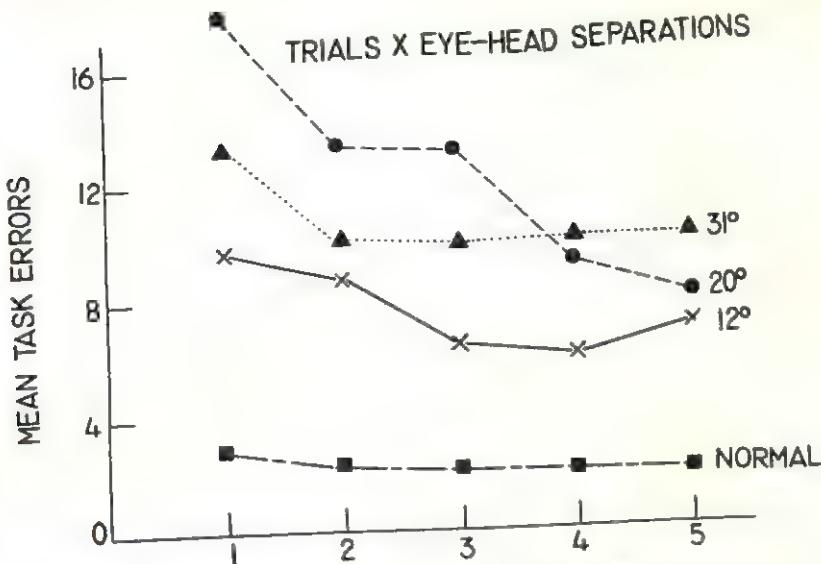


FIG. 5. Driving performance under normal and each of the three eye-head separation conditions with constant visual displacement.

to the right of the driver's line of vision as in the right camera position of Experiment I. Such rotation produced eye-head separation angles of 20°, 12°, and 31° for 0°, 32°, and 51° head displacements.

### Results

The mean performance of all subjects in this experiment under three levels of eye-head separations is shown to be 12.36, 7.72, and 10.74 for the 20°, 12°, and 31° angles. The mean errors were found to be least for the 32° head displacement of a 12° eye-head separation—the smallest angular difference. The mean performance under 0° and 51° head displacements, or 20° and 31° eye-head separations, was not as efficient as that for 32° displacement. Results of an analysis of variance show an overall significant difference for all three eye-head separation conditions ( $F = 29.31, df = 2/22, p < .001$ ). Further examination of the three head displacements with Duncan range tests indicates that 12° separation was significantly different from that of either 20° or 31° angle ( $p < .01$ ). A difference between the latter two conditions was not found.

Driving performance across trials under each of the three eye-head separation conditions, together with performance data from the normal driving condition, is illustrated in Figure 5. Results of the same analysis as in Experiments I and II show an overall

significant difference for the Displacement  $\times$  Trials interaction ( $F = 2.66, df = 8/88, p < .025$ ). The main effects of trials were significant ( $F = 10.86, df = 4/44, p < .001$ ). With the exception of the driving performance under 0° head displacement, which indicated a decreased error over trials, performance under both 32° and 51° head displacement leveled off after the third and second trial, respectively. An obvious fact is that during the last two trials, the driving efficiency under combined conditions was worse than for any of the single conditions in Experiments I or II.

### DISCUSSION

Within the framework of a human driver-vehicle-road tracking system, the automobile was conceived to function as a slave exoskeleton of the driver, who controls and steers the vehicle as an extension of his own body in terms of continuous sensory feedback information from the movement of the vehicle. The longitudinal alignments of vision and head with posture, the vehicle, and the road geometry were assumed to be two of the dynamic referencing mechanisms involved in vehicle control and guidance. These assumptions were experimentally tested in terms of angular displacement of eye and head positions from the longitudinal reference plane in actual driving situations. It was specifically

hypothesized that within a certain tolerance range, angular displacement of eye and head from the longitudinal plane would degrade driver's motor control accuracy and acquisition of vehicle operating skill. Results of the present studies support these assumptions and are discussed in the following sections.

#### *Performance and Angular Displacement*

Previous studies (Smith & Smith, 1962) on the effects of displaced vision on human performance and skill learning confirmed the view that motion is guided and learned in terms of the direction and magnitude of angular displacement from the visual feedback of control movements. Within an indifference range of displacement of visual input, there is no disruption of control movements. Beyond this, angular displacement in a breakdown range disturbs motion coordination and imposes the need for learning in order to achieve effective movement control. The extent of learning or motor refinement required to achieve a given degree of guidance accuracy within the breakdown range varies as a function of the magnitude of displacement. The same assumptions were also applied to head displacement, especially its role in distorting sensory feedback of vehicle movements.

As far as the referencing system is concerned, findings of the present studies and a previous one (Kao & Smith, 1969) suggest that the concepts of angular displacement of vision and head apply as well to the study of vehicular control and guidance. For both the angular displacement of vision with straight head position (Experiment I) and the angular displacement of head with constant straight vision (Experiment II), road tracking performance degraded as a direct function of increasing magnitudes of angular displacement, each at three different levels. As hypothesized, the angular separation of eye and head was found also to degrade performance efficiency as a function of increased magnitude of angular separation when neither was positioned along the longitudinal reference plane (as in Experiment III).

Data from these experiments indicate that from a neurogeometric point of view of sensorimotor control (Gould & Smith, 1963), sensory information from visual feedback and

head displacement may vary in terms of two ranges of tolerance of partial displacement. One range, approximately the left half of the car front in our studies, defines a normal range of displacement for a given car and produces a low level of steering error. Beyond this, the breakdown range approximates the far right of the car front. Displacement between these two loci produces an intermediate level of error, as shown in the center image and center monitor display in Experiments I and II, respectively. Increased angular displacement resulted in magnified steering errors. These findings confirm the predictions made within the framework of a sensory referencing system for the basic control and guidance of vehicles, of which the eye and head are two of the components.

#### *Learning and Angular Displacement*

For the first two experiments, no motor learning was observed in the first three trials under each of the angular displacement conditions. This was found true for Experiment III with combined eye and head displacement. For each level of angular displacement, most learning was complete by the third trial. Performance refinement in road tracking was shown only within the displacement conditions. Performance efficiency was positively related to the minimization of angular or spatial displacement of eye and head, well in accordance with the sensory referencing concept proposed for human vehicular control. This is shown in the first two studies with 0° eye-head alignment and in the third study with 12° eye-head separation producing the least task errors. Such spatial properties of angular displacement are the determining factors in the motor learning process of human vehicle control. Driving learning is the process of minimizing the spatial displacements among vision, head, posture, vehicle, and road geometry through afferent feedback information of vehicle motion in actual operation.

#### *Driver-Generated Motion and Vehicular Guidance*

This research has investigated primarily the sensory aspects of human vehicular control mechanisms in terms of negative feedback

information toward an optimal alignment of the referencing systems for accurate steering. On the motor aspect of the driver-vehicle-road tracking system, the initiation and subsequent precision control of steering performance also depend upon the driver's generation of spatial displacements from the reference plane as the origin of afferent feedback information. This forms the basis of other non-straight control maneuvers such as turning, backing, highway interchange, etc. Continuous steering performance is maintained by so initiating lateral and angular displacement of the eye, the head particularly, and posture in such a way as to align vehicle movement and directional positioning continuously with the new lines of reference.

### *Practical Implications of the Research Findings*

Results of this research help to define the ranges of crucial visual information needed for accurate vehicle control. The most effective visual information is provided by the interactions of the left half of the vehicle hood and the road, with decreasing importance of visual information display as vision is shifted toward the right sectors of the hood. Visual design of the crucial sector of the hood for information display in terms of improved cursor effects could provide a better indication of the vehicle-road dynamic interactions for easy and accurate driving control. A recent study (Kao & Nagamachi, 1969b) showed that night driving accuracy was significantly improved when the outermost corners and center points of the hood were lighted to augment the visual information displayed. This definition of the crucial visual information in a vehicle, together with the basic concept of eye, head, and posture referencing mechanisms of human vehicular control, also has implications in the design of driver trainers and simulators and in the general task specification in driver training. The priority of sensorimotor tasks and information display may be established and presented according to their relative criticality in the vehicle control aspect of the training process.

### REFERENCES

- GORDON, D. A. Perceptual basis of vehicular guidance. *Public Roads*, 1966, 34, 53-68.
- GOULD, J., & SMITH, K. U. Angular displacement of visual feedback in motion and learning. *Perceptual and Motor Skills*, 1963, 17, 699-710.
- HENDRICKSSON, N. G., NILSSON, A., & ANDERSON, A. The driver as receptor of physical impulses. *International Road Safety and Traffic Review*, 1965, 37-40.
- KAO, H. S. R. Feedback concepts of driver behavior and the highway information system. *Accident Analysis and Prevention*, 1969, 1, 65-76.
- KAO, H. S. R., & NAGAMACHI, M. Sensory-motor feedback mechanisms in human vehicular performance. *Ergonomics*, 1969, 12, 741-751. (a)
- KAO, H. S. R., & NAGAMACHI, M. Visual operational feedback and design of vehicle front-end illumination for night driving performance. *Perceptual and Motor Skills*, 1969, 28, 243-246. (b)
- KAO, H. S. R., & SMITH, K. U. Cybernetic television methods applied to feedback analyses of automobile safety. *Nature*, 1969, 222, 299-300.
- KÖHLER, I. Experiments with prolonged optical distortions. *Acta Psychologica*, 1955, 11, 176-178.
- MOSHER, R. S. Exoskeleton prototype. (Technical proposal, Department of the Navy, Office of Naval Research) Schenectady, N. Y.: General Electric Company, 1965.
- MOSHER, R. S., & KNOWLES, W. B. Operator-machine relationships in the manipulator. In, *Human factors of remote handling in advanced systems*. (Tech. Rep. No. 61-430) Wright-Patterson Air Force Base, Ohio: Behavioral Science Laboratory, 1961.
- SMITH, K. U. Review of principles of human factors in design of the exoskeleton and four-legged pedipulator. Madison: University of Wisconsin, Behavioral Cybernetics Laboratory, 1966.
- SMITH, K. U., & SMITH, M. F. *Cybernetic principles of learning and educational design*. New York: Holt, 1966.
- SMITH, K. U., & SMITH, W. M. *Perception and motion*. Philadelphia: Saunders, 1962.
- SMITH, W. M., SMITH, K. U., STANLEY, R., & HARLEY, W. Analysis of performance in televised visual fields. *Perceptual and Motor Skills*, 1956, 6, 195-198.
- STRATTON, G. M. Vision without inversion of the retinal image. *Psychological Review*, 1897, 4, 341-360, 463-481.
- WOOSTER, M. Certain factors in the development of a new spatial coordination. *Psychological Monographs*, 1923, 32 (4, Whole No. 146).

(Received October 18, 1971)

## EVALUATING LANGUAGE TRANSLATIONS: EXPERIMENTS ON THREE ASSESSMENT METHODS<sup>1</sup>

H. WALLACE SINAIKO<sup>2</sup> AND RICHARD W. BRISLIN<sup>3</sup>

*Institute for Defense Analyses, Arlington, Virginia*

Experiments were run to assess three ways of evaluating the quality of language translations: back translation, knowledge testing, and performance testing. Twelve professional English-to-Vietnamese translators processed approximately 10,000 words of technical material (i.e., a helicopter maintenance manual). Subjects took knowledge tests or performed a difficult maintenance task using translated materials. Vietnamese Air Force technicians and U.S. Army technicians served as primary subjects and controls, respectively. The analysis of back translations showed the frequency and types of translation errors that occurred. Knowledge test scores satisfactorily discriminated different quality levels of translations. The performance tests demonstrated (a) the impact of translation quality on performance, (b) the value of working in one's native language (vs. having to learn English), and (c) the importance of providing high-fidelity translations where a complex task is to be done.

Technical documents—maintenance manuals, technical orders, and instructional material—are as critical in the use of complex military equipment as the hardware itself. Training men how to use and service equipment is inevitably tied to the quality of the technical documents they are given. And in the case of material intended for foreign nationals—in this research, the Armed Forces of the Republic of Vietnam—there is an added class of problems: Most of the intended users do not read English, and documents must be translated. In addition, the Vietnamese language contains very few technical terms. Language translation methods are as old as the printed word; but surprisingly, there is almost no literature on the technology of translation and on the accuracy that can be expected from it. One is forced to rely on the subjective views of translators or bilingual readers about the quality of a translated document.

Several experiments were conducted to provide (a) information about different methodologies that could be used to assess the quality of translated technical English and (b) data on factors that affect the quality of text translated from English to Vietnamese. The three assessment techniques examined were back translation, knowledge testing, and performance testing.

### TECHNIQUES

#### *Back Translation*

One method for evaluating translation quality is back translation—specifically, comparing the original English and the back-translated English. In the back-translation technique, the investigator asks one bilingual to translate from the original to the target language, and then he asks another bilingual to translate back from the target to the original. The advantage of the technique is that, as opposed to other methods that have been suggested (e.g., Carroll, 1966; Miller & Beebe-Center, 1956), the translation evaluator does not have to understand or speak the target language. A weakness is the fact that any mistakes in the back translation may be due either to the translator or to the back translator. Thus, even though we evaluate back translation to obtain insights about translation, a perfect translation can be misinterpreted by an incompetent back trans-

<sup>1</sup>The authors would like to thank Vu Tam Ich, Nguyen Nhan, and the officers of Fort Eustis, Virginia, for making this work possible. Further information on all aspects of this investigation (e.g., background, more examples of technical English translated, performance task) can be found in Sinaiko and Brislin (1970).

<sup>2</sup>Requests for reprints should be sent to H. Wallace Sinaiko, who is now at the Smithsonian Institution, Arts and Industries Building, Room 3101, Washington, D. C. 20560.

<sup>3</sup>Now at the Culture Learning Institute, East-West Center, University of Hawaii, Honolulu, Hawaii.

lator, or a good back translator can "correct" a poor translation. This is why back translation should always be complemented by other techniques, such as knowledge testing.

### *Knowledge Testing*

Knowledge testing refers to a method of evaluating translation quality in which subjects read a translated passage and then answer a set of questions about the content of the passage. If subjects can answer all the questions, the translation is assumed to be a good one. While the knowledge-testing technique resembles the standard reading comprehension method, it differs in one important respect: Measures of reading comprehension contain items of graded difficulty and are sensitive to individual differences. Knowledge testing is designed to elicit perfect scores if the translation is good and should be independent of individual differences. The technique was suggested by Miller and Beebe-Center (1956) and by Macnamara (1967) and was first used by Brislin (1970).

This approach asks, "How well can people read and understand Vietnamese that has been translated from English?" The knowledge-testing technique requires the researcher to write a series of questions in English about a passage and then to have them translated. He must also secure subjects who will read the passage and answer the set of questions. Tests must be scored by readers of Vietnamese, too, if they employ fill-in type items. A multiple-choice format obviates the need for a native reader.

### *Performance Testing*

This technique has subjects perform a task requiring them to use either English or translated instructions. To the extent that subjects can complete the task, the translation is regarded as equivalent to the original English text. As in the evaluation techniques previously described, the experimenter does not have to know the target language, since he only has to assess the product of the translated performance instructions.

Performance tests can be scored objectively. In the present experiment, a very demanding 12-step adjustment task on a portion of a

helicopter engine made up the performance test. Three-man crews worked together, and the nature of the task required them to follow written instructions with care. Each of the 12 steps was assessed by a technically qualified observer as "error free," "minor error," or "major error."

Performance testing is the most stringent translation evaluation technique, since it demonstrates the quality of a translation by observable behavior of subjects. However, the technique is the most expensive and time consuming of the three we have used because the experimenter has to (a) define a suitable task, (b) have it translated, (c) provide materials, for example, a helicopter, (d) secure suitably trained subjects, (e) have the subjects perform the task, and (f) obtain the services of observers who are technically competent to grade the task.

## METHOD

### *Bilingual Consultant*

A highly skilled consultant was hired who possessed the following qualifications: Vietnamese native, university teacher in Vietnam, 20 years in the United States, doctoral degree in educational psychology with additional training in linguistics, experience with translating technical materials, and had taught other Vietnamese how to translate.

### *Translators*

A group of 12 bilinguals was hired to provide translation services. At the time of these experiments, 7 of the 12 bilinguals were professional translators. All 12 had worked either part time or full time as translators for an average of 11 years and had translated some technical materials in the past. None, however, had ever translated technical materials as a full-time job.

### *Materials to be Translated*

The 12 bilinguals translated three samples of technical material. The first was a section of the technical manual of the UH-1H helicopter (TM 55-1520-210-20). The second was a set of job performance aids for the C-141A aircraft. More specifically, we used PIMO (Presentation of Information for Maintenance and Operation). These materials have been designed so as to be more understandable than conventional technical manuals. The new format incorporates the following characteristics: organization of tasks based on experimental analysis, a fixed syntax, a standardized verb list, and pictures corresponding closely to printed instructions (Goff, Schlesinger, & Parlog, 1969). The third type of material was the U. S. Air Force's

technical order for the C-141A aircraft (T.O. 1C-141A-2-12). This was chosen so that conventional and job performance aid materials for the same task could be compared.

An example of this material, from Chapter 7 of the UH-1H helicopter manual, is as follows:

7.2. This chapter provides all the instructions and information necessary for maintenance authorized to be performed by organizational maintenance activities on the power train system. The power train is a system of shafts and gear boxes through which the engine drives main rotor, tail rotor, and accessories such as DC generator and hydraulic pump. The system consists of a main drive shaft, a main transmission which includes input and output drives and the main rotor mast, and a series of drive shafts with two gear boxes through which the tail rotor is driven.

Other examples of technical materials translated by the bilinguals can be found in other sections of this article.

### *Translation Tasks*

All 12 bilinguals translated and back translated the three types of technical materials described above for eight hours on 2 different days. For instance, one bilingual would translate on the first day, and another would back translate the first bilingual's work on the second day. All 12 bilinguals worked in quiet rooms and had access to an English dictionary (*Webster's Seventh New Collegiate Dictionary*). The instructions to the subjects were similar to those used by Brislin (1970).

### *Quality Measured by Back Translation*

The efforts of the 12 bilinguals produced 9,558 words of back-translated English, distributed as follows: 2,400 words of the UH-1H technical order, 3,486 words of the C-141A PIMO aids, and 3,672 words of the C-141A technical order.

Every word of the back-translated English was compared to the original, as in the following example: *original English*—Man A performs activity (a test) in flight station; *back-translated English*—Mechanic A carries out the testing while in flight. In this example, the only combination of words that caused an error in the meaning of the back translation as compared to the original English is the substitution of "while in flight" for "flight station." All other words are judged to be equivalent.

The criterion for an error was simply this: Any place in the back translation that is not judged to convey the same meaning as the original English is called a meaning error. Meaning errors could be of six types:

1. An addition—an additional word or phrase appears in the back translation.
2. Minor omission—one or two words from the original are omitted from the back translation.
3. Major omission—same as 2, but involving three or more words.

4. Garbling—three or more words in the back translation are not understandable.

5. Minor substitution—one or two words from the original do not have an equivalent in the back translation, but a phrase replaces the original words (e.g., "flight station" is back translated as "in flight").

6. Major substitution—same as 5, but involving three or more words. Finally, the back translation could be equivalent to the original and marked "O.K."

Our error analysis does not say anything about the operational seriousness of an error. We do not know, for example, whether a substitution error or addition of words would result in poor maintenance to the extent that a helicopter would operate in an unsafe condition.

### *Specific Method of Comparison*

Each of the three types of technical materials (described in Table 1) was arbitrarily divided into phrases averaging from eight to nine words. All phrases either were a complete sentence or contained a complete thought.

Dividing into phrases made it easy to look at a meaningful unit in the original and to find the equivalence or nonequivalence of that unit in the back translation. A given phrase could have more than one error. Each phrase, then, was tallied into one or more of the six error categories, or the "O.K." category. In addition, the exact wording that caused each error was noted.

Since the back translations of all three types of technical materials were examined, comparisons among their error scores can be made. This is possible since either all 12 bilinguals translated and back translated the material (as in the UH-1H technical order) or the 12 bilinguals were randomly assigned to translate or back translate the material (as in the C-141A PIMO aids and C-141A technical order). Thus, the quality of the people involved in work on the three types of material should be equivalent, and any differences should be due to the nature of each type of material. The main back-translation measure was simply a count of the number of meaning errors per passage. A second measure was derived by subdividing the total number of errors into the six categories.

### *Quality Measured by Knowledge Testing*

Two knowledge-testing experiments were run, each using different subjects and materials. In the first experiment, three translations of the same material from the Army's technical manual for the UH-1H helicopter were chosen that were judged to be of different quality. The quality ranking was based on the number of errors in the back translation; that is, Translation A had fewer back-translation errors than Translation B and Translation B had fewer errors than Translation C. In addition, a Vietnamese linguist read the original English and the three translations and then rank ordered the translations from best to worst. His rank ordering was the same as that based on the number of back-translation errors.

The knowledge test consisted of 10 fill-in type questions translated into Vietnamese. The same 10 questions were to be answered after the subject read one of the three translations. Since the questions were the same, any differences in the number answered would be due to the quality of the translations.

Subjects were 68 Vietnamese Air Force enlisted men being trained in helicopter maintenance at Fort Eustis, Virginia. These 68 subjects were randomly assigned to read either translation A ( $n = 22$  men), B ( $n = 23$  men), or C ( $n = 23$  men). Subjects worked in an "open book" mode so that memory was not a factor on this test. An example of a question written about the previously quoted technical passage would be, "Who performs the maintenance on the power train system?" The correct answer is "organizational maintenance."

The second experiment was designed to compare translations of PIMO aids with those for the conventional U. S. Air Force technical order for the C-141A aircraft. A single bilingual translated both the PIMO aids and the technical order. He alternated between sections of one document and the other, so that he would not translate one document better simply because he had practiced on the other.

The questions to be asked about the passages were translated into Vietnamese by the same bilingual. Six of the questions were the same for the technical order and PIMO material, since the same topic was covered in the passages under study. These six questions allowed a range of 0-21 points. The other questions, also representing 21 points, were different for the PIMO and technical order, that is, they were unique to each passage. The "different" questions were added to increase the range of scores. An individual could thus achieve a score of 0-42. The major comparison between the PIMO and technical order would be in the "same" questions, since the same bilingual translated all test materials. Any difference in scores would be in the nature of the PIMO aids or the technical order.

Subjects were 36 Vietnamese Air Force enlisted men being trained in helicopter maintenance at Fort Eustis, Virginia. They read either PIMO or technical order material, and thus there were 18 subjects in a group. These subjects also worked in an "open book" mode. All tests, in both experiments, were scored by a Vietnamese linguist.

### *Quality Measured by Performance Testing*

Although it is a much more expensive and time-consuming approach to evaluating translations, the technique of observing men work with translated material comes closer to an ultimate criterion of the value of translations than any other method: Men do a task that is dependent on written material, and their performance is objectively scored. Good performance means that the writing was accurate and vice versa. In our experiments, teams of technicians carried out a very demanding adjustment task on a portion of the UH-1H helicopter main power plant.<sup>4</sup> Observers, U. S.

Army sergeants who were both experts in helicopter maintenance and instructors on the system to be adjusted, assessed each of 12 steps in the task as "error free," "minor performance error," or "major error." Minor errors were those steps that the crews did wrong but then corrected, major errors were noted if crews could not proceed or if their performance was so poor that it required intervention by the observers.

There were four experimental language conditions: (a) the standard or original English technical manual, (b) a very high-quality translation, and (c) and (d), two lesser grades of translation. The high-quality translation was produced as follows: Two of our best translators each worked independently, then they reviewed each other's work and wrote a "consensus" translation. Finally, our linguist consultant reviewed and modified their combined effort. The translators had available two bilingual glossaries of technical terms. (We refer to this translation as "supervised.")

The first of the lesser quality translations was done by a free-lance, highly qualified translator to whom we gave copies of the same technical glossaries mentioned above. This man worked without review. (We call this the free-lance translation.) The second of the lesser quality translations was obtained by contracting with a Washington, D.C. translation service company for a fixed fee to have the approximately 1,000 words of English translated. We had no control of the method used by the translator nor did he have access to any of our glossaries or other aids. His work also was not reviewed. (We call this the commercial translation.) It is important to note that both the free-lance and commercial translators were highly qualified translators.

Crews used as subjects were assembled from two groups of men at the U. S. Army's Transportation School, Fort Eustis, Virginia: (a) Vietnamese airmen who had just completed the Army's aircraft maintenance and helicopter repairman course and (b) U. S. Army enlisted technicians who were also newly graduated from the same courses. Vietnamese airmen were assigned randomly to one of the four language conditions. In each language condition shown, there were six three-man crews, each of which worked independently. The American Army technicians who used English were tested for comparison purposes.

Only indirect comparisons between the three translation assessment methods can be made since practical considerations made it impossible to test the same materials with the three methods. Brislin (1970) was able to furnish comparative information in an earlier study.

## RESULTS AND DISCUSSION

### *Back Translation*

Reliability of the back-translation examination technique was adequate. Two raters independently examined the 12 back translations for the Army technical manual material, and their ratings of number of errors per passage and types of errors were in close

<sup>4</sup> Section 5-391 "Adjustment—Power Turbine Governor RPM Controls," U. S. Army Technical Manual, TM-55-1520-210-20.

TABLE I  
TRANSLATIONS EVALUATED BY  
KNOWLEDGE TESTING:  
TWO SESSIONS

Translation	No. subjects	Mean score	SD
Session 1: Comparison of three translations of UH-1H technical manual			
A	22	6.1	2.2
B	23	4.3	1.8
C	23	2.6	1.3
Session 2: Comparison of PIMO and technical order translations for C-141A			
PIMO	18		
Total score		34.8	3.3
Same questions		16.2	3.7
Different questions		18.6	1.2
Technical order	18		
Total score		33.2	6.7
Same questions		16.1	2.9
Different questions		17.1	4.7

Note. PIMO = Presentation of information for maintenance and operation.

agreement:  $r = .88$  and  $r = .94$ , respectively. A comparison of the three types of technical material, that is, Army technical manual, Air Force technical order, and job performance aids (PIMO), showed very few differences in types of errors that occurred among translations. The only statistically significant difference was in the proportion of "minor substitution" errors for the Army material (13%) versus both technical order and PIMO material (32% and 30%, respectively). More striking was the fact that for seven categories of error there was very close agreement for translations of all three kinds of material. (A more detailed statement of this error analysis appears in Sinaiko and Brislin, 1970.) The major yield from the back-translation analyses was insight into how the translators went about performing a very difficult task. For example, translators in our experiment did one of four things when they came across unfamiliar words in English or words for which there were no Vietnamese equivalents:

1. They left the English word intact in the translation.

2. They transliterated the word using Vietnamese characters.

3. They coined terms to describe in a functional way the English word or concept. For instance, the translators looked at the word "tachometer" (for which there is no Vietnamese equivalent) and then decided that this meant "rotation measuring device," which they could express. This transformation of difficult technical English to simpler English and then to Vietnamese is called "the explain-around technique" by the present investigators. Wickert (1957) noted that he experienced the same technique when he asked Vietnamese to translate abstract concepts.

### Knowledge Testing

Table I gives the results of both knowledge-testing sessions. For Session 1, where a perfect score is 10, it can be seen that subjects were able to answer more questions about Translation A than B and more about B than C. This rank ordering is the same as that found by errors in the back translation and by the judgments of a Vietnamese linguist. Differences among all combinations of the three means (A versus B, A versus C, B versus C) are statistically significant ( $p < .01$ ). These results show that the knowledge test is sensitive enough to demonstrate differences in translation quality.

For Session 2, the data toward the bottom of Table I show that the translation of the PIMO aids and the technical order for the C-141A allow the same number of both the same (perfect score is 21) and different (perfect score is 21) questions to be answered. Thus, the total number of questions (perfect score is 42) are also the same for the technical order and PIMO aids. The very small differences are not statistically significant ( $p > .10$ ).

### Performance Testing

Table 2 presents the performance results of Vietnamese mechanics working with an English or translated text as well as the results of the control group of U. S. Army mechanics who worked only with an English text. Several striking things about translated technical material are illustrated. First, it is clear that working in one's own language, even if that

material is a translation of a difficult technical manual, is significantly better than having to use a second language. The difference is significant by chi-square at less than the .01 level ( $\chi^2 = 16.6$ ,  $df = 2$ ). However, an important qualification is that the translation must be of high quality. Second, the performance task is sensitive to the quality of translation: Commercial quality<sup>5</sup> produced much higher rates of serious errors than the English text. That is, the Vietnamese airmen worked more effectively with English than they did using a poor translation ( $\chi^2 = 13.5$ ,  $df = 2$ ,  $p < .01$ ). Third, the quality of translated technical documents as measured by performance is significantly influenced by the procedures of the translators. Thus, using a group of men who were approximately equal in their bilingual abilities as translators, we were able to produce very different levels of material. The mode of compensation, that is, placing a premium on speed, was one procedural variable. The availability of bilingual glossaries of technical terms was another.

Incorporation of team translation and a review procedure seemed to make a difference. Finally, the careful translation procedures outlined here can lead to documents that allow Vietnamese mechanics to perform as well as U. S. Army mechanics. (Note, however, that the best Vietnamese groups committed some "major errors," i.e., about 5%, while the Americans did not.)

#### *Subjective Opinions and Translation Quality*

An interesting fact emerged from discussions with some of the Vietnamese airmen who used the best translated material. Most of the men we talked with after they had worked on the performance task expressed a dislike for the translations. The principal objection seemed to be that there were unfamiliar Vietnamese terms used for some of the technical English words. To paraphrase the words of some subjects, "...we did not understand all the Vietnamese words. We would prefer to use the English manual on which we had been trained." It is particularly noteworthy that, in spite of

<sup>5</sup> Data for one crew, commercial translation, were lost because that crew was unable to follow the translation. This supports our contention that this specific translation was poor.

TABLE 2  
PERFORMANCE TEST RESULTS:  
ACCURACY

Experimental condition	% error free	% major errors committed
Vietnamese: Supervised translation	73.1	5.6
Vietnamese: Free-lance translation	40.3	4.2
Vietnamese: Commercial translation	11.0	37.0
English (VNAF subjects)	40.7	20.6
English (U. S. Army subjects)	73.2	0.0

their expressed dislike of even the best quality translation, the measured performance of the airmen was nearly equal to that of the American technicians. At the same time, we asked two bilingual readers (one of whom was an expert in helicopter maintenance) to review and comment on one commercial translation. Each of these men thought that the latter document was "pretty good." However, in practice it resulted in the worst performance of any of the language conditions. The point we wish to underscore is the discrepancy between subjective assessment and performance testing as ways of evaluating translations. The verbal reactions of our subjects and of the linguists were reversed when we actually measured performance.

#### *Recapitulation: Three Methods Compared*

The experiments reported in this study are based on three approaches to assess the quality of translation: (a) back translation, (b) knowledge testing, and (c) performance testing. None of the three methods described requires that the experimenter have proficiency in the target language, although each approach requires the services of linguist translators. Relatively greater demands are placed on translator services in the first two methods than the last; particularly in the use of knowledge testing, translators must be used for the basic English text, the questions to be answered, and as test scorers. Back translation puts an analytic burden on the experimenter that is not present for the other techniques. However, there are no test items to be developed for back translation, while such items are

TABLE 3  
COMPARISON OF THREE TRANSLATION EVALUATION METHODS

Characteristic	Back translation	Knowledge testing	Performance testing
Experimenter proficiency in target language	No	No	No
Translators needed			
Original text	Yes	Yes	Yes
Test items	No	Yes	No
Scoring tests	No	Yes*	No
Back translating	Yes	No	No
Test construction	No	Yes	Yes (but may use available task)
Technical experts as observers	No	No	Yes
Special equipment needed	No	No	Yes
Relative cost; face validity	Lowest	Middle	Highest
Confidence in results	Lowest	Middle	Highest
Test subjects	No	Yes—any reader of the language	Yes—must be trained in the task

\* If fill-in type items are used.

at the heart of knowledge tests. Performance testing may require that a task be designed, although, as in the present experiments, an available task was used. In addition to translators, test subjects are required for knowledge and performance testing; this is not so with back translation. Only in the case of performance tests are technical experts needed to evaluate what subjects do. Similarly, special equipment or material is needed for performance tests but not for the other two approaches (see Table 3). The relative costs of the three methods are probably in this order (low to high): back translation, knowledge tests, and performance tests. Finally, confidence in results or face validity of the methods is likely in the same order.

#### REFERENCES

BRISLIN, R. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1970, 1, 185-216.

CARROLL, J. Quelques mesures subjectives en psychologie: Fréquence des mots, significativité et qualité de traduction. *Bulletin de Psychologie*, 1966, 19, 580-592.

GOFF, J., SCHLESINGER, R., & PARLOG, J. *PIMO test summary*. (Tech. Rep. 69-155, Vol. II) Andrews Air Force Base, Md.: Space and Missile Systems Organization, Air Force Systems Command, May 1969.

MACNAMARA, J. The bilingual's linguistic performance—a psychological overview. *The Journal of Social Issues*, 1967, 23, 58-77.

MILLER, G. A., & Beebe-Center, J. Some psychological methods for evaluating the quality of translation. *Mechanical Translation*, 1956, 3, 73-80.

SINAICO, H. W., & BRISLIN, R. W. Experiments in language translation: Technical English-to-Vietnamese. (Research Paper P-634) Arlington, Va.: Institute for Defense Analyses, 1970.

SINAICO, H. W., Guthrie, G. M., & Abbott, P. S. *Operating and maintaining complex military equipment: A study of training problems in the Republic of Vietnam*. (Research Rep. P-501) Arlington, Va.: Institute for Defense Analyses, 1969.

WICKERT, F. An adventure in psychological testing abroad. *American Psychologist*, 1957, 5, 86-88.

(Received December 14, 1971)

## THE RELATIONSHIP BETWEEN CONSUMERS' CATEGORY WIDTH AND TRIAL OF NEW PRODUCTS

JAMES H. DONNELLY, JR.<sup>1</sup> AND MICHAEL J. ETZEL

*University of Kentucky*

SCOTT ROETH

*IBM Corporation, Lexington, Kentucky*

The study tested the hypothesis that individuals who exhibit broad "category ranges" in judging stimuli will be more apt to try new products. A sample of housewives completed Pettigrew's Category Width Scale, and responded to questions about their trial of five new grocery products. It was found that consumers' tolerance for errors of exclusion and inclusion was related to whether or not they purchased the products.

A consumer's willingness to try new products has been related to several behavioral and demographic dimensions. For example, housewives who acted as clothing innovators had better educations, had higher incomes, had husbands with higher occupational status, had been exposed to more magazines, and had been involved with more organizations than non-innovators (Summers, 1970). Additional research has shown that early product tryers are highly mobile socially and geographically, have higher educational levels (Opinion Research Corporation, 1959), have strong needs for achievement, dominance, change, and exploration (Won't they try new products, 1959), are very realistic in their aspirations (White, 1966), are inner-directed (Donnelly, 1970), and are low in dogmatism (Blake, Perloff, & Heslin, 1970; Jacoby, 1971).

In order to further expand the profile of early tryers, the present study examined the relationship between a housewife's purchase of new food products and her acceptance of qualitatively different forms of risk. This relationship has not previously been studied in an actual buying situation. The initial work in the area (Popielarz, 1967) investigated willingness to try new products in hypothetical purchase situations.

In applying the risk-taking model, it appears reasonable that the purchase of new products can be a potentially high-risk situa-

tion, since the new product may be relatively unfamiliar. However, this study follows the ideas of Pettigrew (1958) and Popielarz (1967) that decisions to try new products may actually present the decision maker with two kinds of risk. Specifically, the consumer who tries new products appears to be willing to risk purchasing some products with which he may not be satisfied, while the individual who restricts his purchases to products and/or brands with which he is familiar is rarely dissatisfied with inferior products. However, the latter runs the risk of avoiding many products that would provide him with more satisfaction than the ones he currently uses.

An analogy to statistical decision making can be drawn from the above situations. When making the decision, the individual establishes levels of probability beyond which he is not willing to commit errors of (a) rejecting a hypothesis as being false when it is actually true and (b) accepting a hypothesis as being true when it is actually false. What the decision maker is actually doing is setting his Type I and Type II error limits. In the context of buyer behavior the consumer willing to try new products would have a high tolerance for making Type I errors or errors of *inclusion*. When he tries new products he tends to accept them as satisfactory when they may, upon use, prove to be unsatisfactory. Since he may not "reject" certain products, he risks including some inferior or unsatisfactory products in his selection. On the other hand, the nontryer would have a high tolerance for making Type II errors or errors

<sup>1</sup> Requests for reprints should be sent to James H. Donnelly, Jr., Department of Business Administration, University of Kentucky, Lexington, Kentucky 40506.

of *exclusion*. Since he relies on tried and proven products, he risks not trying some new products which would actually benefit his selection (Kogan & Wallach, 1964; Popielarz, 1967).

Since perceived risk is a cognitive phenomenon, psychological research on cognition seems to offer a means for expanding the consumer risk-taking proposition. One aspect of cognitive style which appears particularly applicable to new product acceptance is the concept of category width investigated by Pettigrew (1958). It has been found that individuals reveal marked consistency in the category widths they perceive relative to various stimuli (Bruner, Goodnow, & Austin, 1956). That is, in selecting maximum values of various optical and auditory phenomena, subjects consistently had either a broad, medium, or narrow range of judgment. An individual labeled as a broad categorizer tended to judge extreme instances of a category more distant from a central tendency value relative to the judgments of an individual labeled as a narrow categorizer. Pettigrew (1958) offered a possible explanation for the relationship between category width and tolerance for errors of exclusion and inclusion:

Broad categorizers seem to have a tolerance for Type I errors; they risk negative instances in an effort to include maximum positive instances. By contrast, narrow categorizers are willing to make Type II errors. They exclude many positive instances by restricting their category ranges in order to minimize the number of negative instances [p. 532].

The present study explores the relationship between a consumer's tolerance for errors of exclusion and inclusion and the purchasing of new products. Specifically, the hypothesis tested was that an individual's breadth of categorization is related to his willingness to buy new products. That is, individuals with broad category ranges will be more apt to buy new products than individuals with narrow category ranges.

## METHOD

### Procedure

Five food product groups were selected for analysis. In selecting these groups, the researchers did an extensive review of new food products in order to develop a group of products that differed significantly

from anything previously available. In rating the newly available products on their degree of departure from what was already on the market, the researchers used four product characteristics—packaging, physical appearance, required user preparation, and manufacturer's technological processing. This was done to insure that the products used in the study would be noticeably different from older products already on the market and to account for personal experience with both the product (seeing it in a store) and with manufacturer's advertising. For the hypothesis to be accepted, therefore, more broad categorizers would have to purchase the products than narrow categorizers.

Using these four characteristics, the researchers selected five product groups for the study: frozen pudding, frozen donuts, canned pudding, canned cake frosting, and instant oatmeal with fruit. Canned pudding, for example, differs in packaging, user preparation, physical appearance, and technological processing from the established competing products.

### Subjects

The subjects were 175 randomly selected housewives in central Kentucky who indicated familiarity with the five products. They were the remainder of an original sample of 210 which was reduced when subjects who exhibited a lack of familiarity with some or all of the products were dropped. Each housewife, personally interviewed during the spring of 1970, was asked which of the products she had purchased.

To determine the subjects' breadth of categorization, the researchers utilized Pettigrew's (1958) Category Width Scale. In each question of the 20-item instrument the respondent is given a hypothetical average value for some series of events. Four alternative responses are provided for each of the maximum and minimum estimates, each of which is weighted as a function of the extent to which it deviates from the average value for the item. Following is an example of a test item (the numbers in parentheses are scoring weights):

For the past twenty years, Alaska's population has increased an average of 3,210 people per year. What do you think:

- a. was the *greatest* increase in Alaska's population in a single year during these twenty years?
  - 1. 6,300 (2)
  - 2. 21,500 (3)
  - 3. 3,900 (0)
  - 4. 4,800 (1)
  
- b. was the *smallest* increase in Alaska's population in a single year during these twenty years?
  - 1. 470 (3)
  - 2. 1,960 (1)
  - 3. 980 (2)
  - 4. 2,520 (0)

Each of the subjects completed the Category Width Scale. A respondent's score was derived by summing the weights of the chosen alternatives for all items. The sum of the distance scores (0-3 on each item) is called the respondent's category width. High

scores indicate broad categorizers (high tolerance for errors of inclusion) and lower scores indicate narrow categorizers (high tolerance for errors of exclusion). The range of possible scores on the Category Width Scale is from 0 to 120.

### Analysis

Since fifteen of the 175 respondents failed to complete the Category Width Scale, they were deleted from the sample. This resulted in a total of 160 usable responses. The sample was divided into thirds ( $n = 53, 54, 53$ ) according to the scale scores. In accord with Pettigrew's (1958) instructions, the individuals in the highest third were classified as broad categorizers, while those in the lowest third were classified as narrow categorizers. The total of 53 broad and 53 narrow categorizers ( $n = 106$ ) constitute the data presented here.

The chi-square test was used in order to arrive at the probability levels which distinguish broad and narrow category width results in relationship to the purchase of new food products. The use of this test provided a measure of the independence of the two sets of classification results: broad and narrow categorizers and purchase of new food products. The results are presented in Table 1.

### RESULTS

The data were analyzed to determine if a relationship existed between a housewife's breadth of categorization and her willingness to purchase new food products. Out of the five products tested, four chi-square probability levels below .05 were produced.

### DISCUSSION

The results of the study support the hypothesis that an individual's breadth of categorization is related to the purchase of new products. It appears that one's trial of new products involves a propensity to assume different kinds of risk. Specifically, a willingness to try genuinely new products seems to be associated with a tolerance for errors of inclusion, while unwillingness to try them involves a tolerance for errors of exclusion.

The findings of this study raise important questions for marketers covering the similarity or dissimilarity of new product introductions from previously available products, since a new product that is too dissimilar may preclude a certain segment of the market (narrow categorizers) from trying it.

The study concentrated on broad categorizers and their acceptance of new products.

TABLE 1  
CHI-SQUARE PROBABILITY LEVELS OF BREADTH OF CATEGORIZATION AND PURCHASE OF NEW FOOD PRODUCTS

Product	Had tried	Had not tried	$\chi^2$ probability level
Frozen pudding			<.01
Broad categorizers	17	36	
Narrow categorizers	4	49	
	5	5	
Frozen donuts			<.02
Broad categorizers	26	27	
Narrow categorizers	14	39	
Canned pudding			<.02
Broad categorizers	27	26	
Narrow categorizers	15	38	
Canned cake frosting			<.02
Broad categorizers	37	16	
Narrow categorizers	24	29	
Instant oatmeal with fruit			<.25
Broad categorizers	35	18	
Narrow categorizers	29	24	

Note.  $N = 106$ .

This does not mean, however, that narrow categorizers should not be of interest to researchers in the area of new product acceptance. For example, a study in which new products are selected on their degree of similarity (rather than dissimilarity) from what was previously available may show very different results from the ones reported here. In fact, narrow categorizers may show a purchase rate greater than, or equal to, broad categorizers. Thus, we may add a whole new dimension to the study of the diffusion of new products by finding that product attributes are as important as behavioral and demographic characteristics in identifying the innovators for particular products.

### REFERENCES

- BLAKE, B., PERLOFF, R., & HESLIN, R. Dogmatism and acceptance of new products. *Journal of Marketing Research*, 1970, 7, 483-486.
- BRUNER, J., GOODNOW, J. J., & AUSTIN, G. *A Study of Thinking*. New York: Wiley, 1956.
- DONNELLY, J. Social character and acceptance of new products. *Journal of Marketing Research*, 1970, 7, 111-113.
- JACOBY, J. Personality and innovation proneness. *Journal of Marketing Research*, 1971, 8, 244-247.
- KOGAN, N., & WALLACH, M. A. *Risk Taking*. New York: Holt, Rinehart & Winston, 1964.

- OPINION RESEARCH CORPORATION, *America's Taste-makers*. (Research Report No. 1) Princeton: Author, 1959.
- PETTIGREW, T., The measurement and correlates of category width as a cognitive variable. *Journal of Personality*, 1958, 26, 532-544.
- POPIELARZ, D., An exploration of perceived risk and willingness to try new products. *Journal of Marketing Research*, 1967, 4, 368-372.
- SUMMERS, J., The identity of women's clothing fashion opinion leaders. *Journal of Marketing Research*, 1970, 7, 178-185.
- WHITE, I. The perception of value in products. In J. W. Newman (Ed.), *On knowing the consumer*. New York: Wiley, 1966.
- Won't they try new products? They're not willful. *Advertising Age*, April 13, 1959, 3.

(Received October 27, 1971)

---

Manuscripts Accepted for Publication in the  
*Journal of Applied Psychology*

- Employee Reactions to a Pay Incentive Plan. Cortlandt Cammann (Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan 48106) and Edward E. Lawler III.
- Measurement of Attitudes of Industrial Work Groups toward Psychology and Testing. Rodney F. Roads and Frank J. Landy (Department of Psychology, Pennsylvania State University, 417 Psychology Building, University Park, Pennsylvania 16802).
- Effects of Conflict Management Training upon Police Performance. Joseph Zacker (Department of Psychology, City College, City University of New York, 138th Street and Convent Avenue, New York, New York 10031) and Morton Bard.
- Cross-Cultural Differences in Two-Factor Motivation. George H. Hines (Senior Lecturer, Victoria University of Wellington, P. O. Box 196, Wellington, New Zealand).
- Prediction of Advanced Level Aviation Performance Criteria from Early Training and Selection Variables. Ronald M. Bale, George M. Rickus, Jr., and Rosalie K. Ambler (Naval Aerospace Medical Research Laboratory, Naval Aerospace Medical Institute, Naval Aerospace Medical Center, Pensacola, Florida 32512).
- A Study of the Relationship between Need Satisfaction and Absenteeism among Managerial Personnel. Lawrence G. Hrebiniak (College of Business Administration, Pennsylvania State University, 211-B Boucke Building, University Park, Pennsylvania 16802) and Michael R. Roteman.
- Likert Organizational Profile: Methodological Analysis and Test of System 4 Theory in Brazil. D. Anthony Butterfield (Alfred P. Sloan School of Management, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, Massachusetts 02139) and George F. Farris.
- Race of Employment Interviewer and Reasons Given by Black Job Seekers for Leaving their Jobs. James Ledvinka (Assistant Professor, Department of Management, College of Business Administration, University of Georgia, Athens, Georgia 30601).
- A Preliminary Study of the Effects of Crash Helmet Visor Color on Color Recognition. Harold D. Warner (Assistant Professor of Psychology, Department of Social Sciences, University of Missouri, Rolla, Missouri 65401).
- Response Distortion on the GPI and GPP in a Selection Context: Some Implications for Predicting Employee Tenure. Donald P. Schwab (College of Business and Economics, University of Kentucky, Lexington, Kentucky 40506) and Gary L. Packard.
- Relationship of Information Processing Attitude Structures to Private Brand Purchasing Behavior. James R. Bettman (Graduate School of Business Administration, University of California, Los Angeles, California 90024).
- An Automated Patient Behavior Checklist. Donald W. Morgan, Jeffrey L. Crawford, Sinai I. Frenkel (Deputy Director, COMPSY, Department of Psychiatry and Neurology, Walter Reed General Hospital, Washington, D.C. 20012), and James L. Hedlund.
- Effects of Curtailment on an Admissions Model for a Graduate Management Program. V. Srinivasan and Alan G. Weinstein (Graduate School of Industrial Administration, Carnegie-Mellon University, Schenley Park, Pittsburgh, Pennsylvania 15213).
- Job Interview Training with Rehabilitation Clients: A comparison of Videotape and Role-playing Procedures. Marlene G. Venardos (3408 Calle del Ranchero N.E., Albuquerque, New Mexico 87106) and Mary B. Harris.
- Impact of Alternative Compensation Systems on Pay Valence and Instrumentality Perceptions. Donald P. Schwab (College of Business and Economics, University of Kentucky, Lexington, Kentucky 40506).
- Effects of Goal Setting and Supervision on Worker Behavior in an Industrial Situation. W. W. Ronan, Gary P. Latham (School of Psychology, University of Akron, Akron, Ohio 44304), and S. B. Kinne, III.
- Relation of Price to Perception of Quality of New Products. Arch G. Woodside (College of Business Administration, University of South Carolina, Columbia, South Carolina 29208).
- Facilitation of Learning from a Technical Manual: An Exploratory Investigation. Neil C. Kalt (Department of Psychology, Baruch College, City University of New York, 17 Lexington Avenue, New York, New York 10010), and Katherine Merlo Barrett.
- Influence of Sex-Role Stereotypes on Personal Decisions. Benson Rosen (Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514) and Thomas H. Jerdee.
- Prediction of Artistic Performance from Biographical Data. Lawrence R. James (NRC Research Associate, Department of the Navy, Navy Medical Neuropsychiatric Research Unit, San Diego, California 92152), Robert L. Ellison, David G. Fox, and Calvin W. Taylor.

## SHORT NOTES

### CAREER ORIENTATION AND JOB SATISFACTION AMONG WORKING WIVES

MARTIN J. GANNON<sup>1</sup> AND D. HUNT HENDRICKSON

*Department of Business Administration, University of Maryland*

Two aspects of "career orientation" are shown to be factorially independent of one another among 69 working wives. These aspects appear to be "job involvement" and "the relative importance of work over the family." While job involvement was shown to be related positively and significantly to job satisfaction, the relative importance of work over the family was not.

Several social scientists have emphasized the concept of career orientation in the explanation of organizational efficiency and effectiveness (see Goode, 1960; Weber, 1947). Operationally, the major studies of career orientation have focused on job involvement, which has been shown to be related to such criteria as employee turnover and absenteeism (Weissenberg & Gruenfeld, 1968). Generally speaking, job involvement refers to the commitment of the individual to his work (Lodahl & Kejner, 1965).

Although career orientation has been of major interest to social scientists, its importance among working wives has received minimal attention. Such inattention persists even though women constitute the most dynamic element in the growth of the labor force. From March 1971 to March 1972, the number of working wives increased by 719,000 to 19.3 million (*Manpower Report of the President*, Table B-1, 1973). In addition to this influx of wives into the labor market, a new militancy has arisen that has been critical of the work roles that females presently fulfill in organizations. For at least these two reasons, a study of the relationship between career orientation and job satisfaction among working wives seems appropriate.

#### METHOD

The study was conducted among retail employees who were either clerks or office workers in six establishments located in the Washington, D.C. area. Only working wives were included in the sample. Of the 126 questionnaires that were distributed, 77 were returned and 69 were usable. The usable response rate was 56%.

Job satisfaction was measured through the use of the Cornell Job Description Index (JDI). Career

<sup>1</sup> The authors wish to thank Stephen J. Carroll, Jr. for helpful criticisms and suggestions.

Requests for reprints should be sent to Martin J. Gannon, College of Business Administration, University of Maryland, College Park, Maryland 20742.

orientation was investigated through the use of seven items constructed in terms of a 5-point scale of disagreement-agreement. These questions were selected after an investigation was completed of the job-involvement scale developed by Loda and Kejner (1965). However, the present researchers decided to focus on the broader concept of career orientation rather than merely on job involvement. Further, the researchers, in conjunction with the managers in the six establishments under study, wanted to utilize specific items that seemed to be germane to the particular population (working wives).

The seven items concerned with career orientation were then factor analyzed (principal component method, orthogonal rotation, BMD 03M). The matrix was rotated with the eigenvalue specified at 1.

#### RESULTS

Two factors emerged from the orthogonal rotation. As shown in Table 1, four items were loaded

TABLE 1  
ROTATED LOADINGS ON FACTORS CONCERNED  
WITH CAREER ORIENTATION (N = 69)

Item	Factor	Loading
I would be willing to work overtime if my boss asked me	.78	.07
I regard my job as important to me as my family	.17	.71
I want to know about other phases of the business besides the one in which I am employed	.06	.18
I would continue to work if I were given 100,000 dollars	.60	.13
Being satisfied with my job is important for my overall satisfaction	.66	.07
I would come to work if my 9-year-old son were home from school sick with a cold	-.01	.82
I would be willing to travel occasionally overnight for my job	.62	.17

TABLE 2  
CORRELATIONS BETWEEN CAREER ORIENTATION AND  
JOB SATISFACTION ( $N = 69$  WORKING WIVES)

Indices of job satisfaction (JDI)	Job involvement (Factor 1)	Work versus family orientation (Factor 2)
Work	.31**	.22
Supervision	.30**	.17
People	.27 *	.06
Pay	.17	.05
Promotion	.16	.27*
Total	.36**	.22

\*  $p \leq .05$ .

\*\*  $p \leq .01$ .

heavily on the first factor, and all of them pertain to job involvement. The second factor was loaded heavily by two items, and both appear to concern the relative importance of work over family activities. The two items loaded heavily on the second factor were (a) I regard my job as important to me as my family; and (b) I would come to work if my 9-year-old son were home from school sick with a cold.

As shown in Table 2, job involvement was significantly and positively correlated with the overall score of the JDI. In addition, job involvement was associated significantly with three of the five subscales of the JDI: work, supervision, and people. Conversely, the relative importance of work over family activities was not significantly related to the overall score of the JDI. It was, however, associated with one subscale: As the relative importance of work over family activities increased, satisfaction with promotion rose. Obviously it is possible that a successful rate of promotion in the past contributes to a higher degree of importance of work relative to family activities rather than vice versa. However, the relative importance of work over family activities was not correlated with any of the other subscales of the JDI.

#### DISCUSSION

Similar to the sample of males studied by Weissenberg and Gruenfeld (1968), the working

wives in this study were more satisfied with their jobs when job involvement was high. This finding suggests that the influence of job involvement on job satisfaction is similar among males and females.

The study also showed that job involvement was factorially independent of the relative degree of concern for work versus the family. Thus, working wives with a strong family orientation were just as likely to be committed to the job as those working wives with a relatively smaller degree of family orientation. In addition, job involvement itself had a stronger relationship to job satisfaction than the work versus the family orientation. These findings may indicate that working wives are simultaneously capable of showing high interest and concern both for the job and the family or, in other words, have a form of "dual allegiance" to the family and the job that many male workers have both to the company and the union (England, 1960). Hence, discussions of working wives in terms of a work versus family dichotomy appear to be quite oversimplified.

While the present study has been exploratory, it has suggested that job commitment is related to job satisfaction in a similar fashion among males and females, that it is factorially independent of the issue of family versus work, and that the dual allegiance of females to the family and work does not appear to influence job satisfaction.

#### REFERENCES

- ENGLAND, G. Dual allegiance to company and union. *Personnel Administration*, 1960, 23, 20-25.
- GOODE, W. Encroachment, charlatanism, and the emerging profession: Psychology, sociology, and medicine. *American Sociological Review*, 1960, 25, 902-914.
- LODAHL, T., & KEJNER, M. The definition and measurement of job involvement. *Journal of Applied Psychology*, 1965, 49, 24-33.
- Manpower Report of the President. Washington, D.C.: U.S. Government Printing Office, 1973.
- WEBER, M. *The theory of social and economic organization*. Glencoe: Free Press, 1947.
- WEISSENBERG, P., & GRUENFELD, L. Relationship between job satisfaction and job involvement. *Journal of Applied Psychology*, 1968, 52, 469-473.

(Received September 14, 1971)

## JOB ATTITUDES AS PREDICTORS OF TERMINATION AND ABSENTEEISM: CONSISTENCY OVER TIME AND ACROSS ORGANIZATIONAL UNITS

L. K. WATERS<sup>1</sup>

*Ohio University*

DARRELL ROACH

*Nationwide Insurance, Columbus, Ohio*

Two replications of a study by Waters and Roach (1971) concerned with job satisfaction measures as predictors of withdrawal behavior were conducted. Only three variables (two concerned with the work itself and an overall job satisfaction rating) were consistent predictors of both permanent and temporary withdrawal from the work situation.

Numerous studies have focused on the attitudinal correlates of various forms of withdrawal from the work situation. Reviews of the literature by Brayfield and Crockett (1955), Herzberg, Mausner, Peterson, and Capwell (1957), and Vroom (1964) have indicated some consistency in reported relationships between job-related attitudes and both termination and absenteeism criteria. However, there has been little evidence concerning the replication of predictions of withdrawal behavior either across organizational units or over time periods within a single unit of the organization. The purpose of the present study was to replicate a recent study by Waters and Roach (1971) in a second regional office of a national insurance company and over a second-year time period within one regional office.

### METHOD

In order to replicate the original study, data were obtained for two samples of female clerical workers. One sample consisted of 80 workers at one regional office of a national insurance company who had remained on the job for more than 1 year after administration of a job attitude questionnaire (see Waters & Roach, 1971, for a report of a 1-year follow-up for this regional office). Of the 80 workers, 62 were still employed at the end of 2 years, and 18 had terminated during the second year for reasons other than pregnancy or retirement at age 65. Six employees had transferred to other sections of the company, and no records were available. For the 62 current employees, frequency of absence data for the second-year period were obtained from company records.

The second sample consisted of 117 workers at a second regional office of the same insurance company. The second regional office was located in the same

state and in a somewhat larger metropolitan area. The 117 women included 90 who were still with the company 1 year after administration of a job attitude questionnaire and 27 who had terminated for reasons other than pregnancy or retirement at age 65. For the 90 workers who were still on the job at the end of 1 year, frequency of absence data were obtained from company records.

The job attitude questionnaire had been administered to both samples at their place of work. The job attitude scales were presented in booklet form and consisted of separate overall satisfaction and dissatisfaction scales (always the first two scales, order randomized), a bipolar satisfaction/dissatisfaction scale, the five scales of the Job Description Index (Smith, Kendall, & Hulin, 1969), and a list of 11 job factors (arranged in alphabetical order) to be rated on a satisfaction/dissatisfaction scale. Ratings of satisfaction/dissatisfaction (both for overall and specific job factors) were made on a 12-point bipolar scale and the separate overall satisfaction and dissatisfaction ratings on 7-point scales, which consisted of the appropriate 6 points of the 12-point satisfaction/dissatisfaction scale plus a seventh alternative (not satisfied or not dissatisfied).

### RESULTS AND DISCUSSION

In the original study, satisfaction with five of the seven "intrinsic" variables and two overall job attitude measures correlated significantly ( $p < .05$ ), but of relatively low magnitude, with the termination criterion. Of these variables, four showed consistently significant relationships both over time and across regional offices. The four variables (with correlations listed for the original 1-year follow-up, the 1-year follow-up at the second regional office, and 1- to 2-year follow-up at the original office) were the Work itself-Likert-type scale (.22, .28, .26), the JDI Work scale (.24, .26, .40), the overall satisfaction rating (.23, .27, .22), and the bipolar overall satisfaction/dissatisfaction rating (.24, .27, .22). No additional variables cross-validated from the 1-year to the 2-year follow-up at the original regional office, and

<sup>1</sup> Requests for reprints should be sent to L. K. Waters, Department of Psychology, Ohio University, Porter Hall, Athens, Ohio 45701.

the only other variable that was a consistent predictor across regional offices was the Likert-type item dealing with responsibility on the job (.38 at the original and .19 at the second regional office).

Although 9 of the 22 variables in the original study were significantly ( $p < .05$ ) related to frequency of absences, the relationships were low and no observable pattern of predictors was evident. Four of the nine variables were consistent predictors of absenteeism in both replications. Three of these were the same variables that replicated in predicting permanent withdrawal behavior, termination (with correlations listed for the original office, the second regional office, and the 1- to 2-year follow-up at the original office): Work itself-Likert-type item (-.20, -.40, -.32), the JDI Work scale (-.28, -.34, -.34), and overall satisfaction (-.28, -.38, -.34). Also, job level in the organization (-.23, -.37, -.26) replicated over time and across regional offices. The negative relationships resulted from fewer absences being associated with higher levels of satisfaction or job position. Four other variables showed low but significant correlation with absenteeism across the two regional offices (two Likert-type items dealing with salary and sense of achievement, the bipolar satisfaction/dissatisfaction scale, and years with the company), but no other variables cross-validated from the 1-year to 2-year follow-up at the original regional office). For those employees who remained on the job over the 2-year period at the same regional office, the correlation between frequency of absences for the first and second year was .55.

In this study, for a given sample, several predictors were found to be related to one or the other type of withdrawal behavior.

However, considering replications across both time and organizational units, only three variables (two dealing with the work itself and an overall satisfaction measure) were consistent predictors of both permanent and temporary forms of withdrawal from the work situation. These data suggested that studies covering one time period at one organizational unit may overemphasize the saliency of satisfaction variables in predicting termination and absenteeism. While a limited number of satisfaction variables were found in this study to be consistent predictors of two types of withdrawal behavior, the magnitude of their correlations was of questionable practical significance. Further efforts to determine attitudinal predictors of withdrawal behavior without replication or consideration of situational and personal variables seem of rather limited value.

#### REFERENCES

- BRAYFIELD, A. H., & CROCKETT, W. H. Employee attitudes and employee performance. *Psychological Bulletin*, 1955, 52, 396-424.
- HERZBERG, F., MAUSNER, B., PETERSON, R., & CAPWELL, D. *Job attitudes: Review of research and opinion*. Pittsburgh: Psychological Service of Pittsburgh, 1957.
- SMITH, P. C., KENDALL, L. M., & HULIN, C. L. *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes*. Chicago: Rand McNally, 1969.
- VROOM, V. *Work and motivation*. New York: Wiley, 1964.
- WATERS, L. K., & ROACH, D. Relationship between job attitudes and two forms of withdrawal from the work situation. *Journal of Applied Psychology*, 1971, 55, 92-94.

(Received December 31, 1971)

## INTERVIEW DECISIONS AS DETERMINED BY COMPETENCY AND ATTITUDE SIMILARITY

GLEN D. BASKETT<sup>1</sup>

*Georgia Institute of Technology*

Fifty-one subjects were asked to assume that they worked for a large company and that the president had asked them to evaluate a candidate for a position as a vice president. The target's dossier and information concerning 10 of his attitudes were given to the subjects as stimuli for the evaluation. Three levels of the target's competency and two levels of attitude similarity between the subject and the target were varied in a  $3 \times 2$  factorial design to examine their effects on subsequent job recommendations and suggested salaries. Similarity tended to influence the recommendation ( $p < .10$ ) and did significantly influence salary ( $p < .05$ ). Competency was shown to influence both the recommendation ( $p < .001$ ) and salary ( $p < .001$ ). The implications for industry are discussed.

Since one of the major factors in the decision to hire an applicant is the outcome of the employment interview, it is understandable that recent experimental attention has been devoted to it. Among the findings of this research are that negative information apparently receives greater weight in the decision than does positive information (Carlson, 1967; Mayfield & Carlson, 1966; Miller & Rowe, 1967; Springbett, 1958); interviewers seem to seek out negative information (Bolster & Springbett, 1961; Springbett, 1958); the favorability-unfavorability of written information is more important than the applicant's photograph (Carlson, 1967); and the order of presentation of favorable-unfavorable information is important (Bolster & Springbett, 1961).

These studies have generally relied on the use of one or more standard "target" applicants in an attempt to ascertain the properties of the stimuli that elicit the final decision. However, they have overlooked at least one aspect of the target that might have relevance to the interviewer; i.e., how much the applicant agrees with the interviewer (Griffitt & Jackson, 1970).<sup>2</sup>

It has been stated that in the interview process the judge may be seeking negative information about the applicant (Bolster & Springbett, 1961). Clearly, his task is to make an evaluation concerning the applicant—to hire or not to hire—and it is reasonable for him to use all the obtainable information that he feels is relevant to the decision. Could it be that information is also

gathered concerning the applicant's attitudes and that the extent of agreement with the judge is also influencing the decision?

There is a substantial body of research in interpersonal attraction which might be relevant to this situation. Byrne and his coworkers have conducted a number of studies over the last decade that indicate that a certain class of evaluations are influenced by the extent of agreement between the judge and the target (Byrne, 1961, 1969; Byrne, Baskett, & Hodges, 1971; Byrne & Clore, 1970; Griffitt & Jackson, 1970). These results have been cast within a reinforcement framework that postulates that an agreement provides consensual validation for the judge's beliefs. This consensual validation acts as a reinforcer by eliciting the positive affect associated with the target. Disagreement elicits negative affect, and the combination of affect is reflected in the evaluation of the target. On the basis of this research one would predict that when a judge is presented a target similar to himself, he will evaluate the target more favorably than if the target had been dissimilar and would, therefore, be more likely to recommend hiring the target. Additionally, one might predict that a second but related decision concerning the amount of salary offered to a candidate is also positively influenced by the amount of similarity-dissimilarity experienced (Griffitt & Jackson, 1970).

It was assumed that the more competent the candidate for a job, the more valuable he should be, and thus more competent candidates should receive stronger recommendations and be offered a higher salary (Griffitt & Jackson, 1970).

The present experiment was designed to test these predictions: First, the more similar the target is to the judge, the stronger the recom-

<sup>1</sup> Requests for reprints should be sent to Glen D. Baskett, Department of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332.

<sup>2</sup> However, Syliaha (1962) reported on the comparison of the replies of candidates with the judge's replies on a self-description form and found inconsistent results for predicting and acceptance or rejection of the candidate.

mendation and the higher the salary offered; second, the more competent the target, the stronger his recommendation and the more money offered as salary.

### METHOD

Fifty-one college students participated as subjects out of 62 who were asked to volunteer. The subjects were recruited from two social psychology classes at Georgia Institute of Technology and received additional grade credit for participation. Forty-eight subjects were males.

As part of a class assignment, the students responded on a 53-item attitude questionnaire. Ten of these attitudes were either modified Interpersonal Trust statements (Rotter, 1967) or statements reflecting the subject's faith in the environmental protection process. All 10 statements were answered on a 7-point Likert-type scale. Each subject's responses to these 10 statements were used to prepare a target person who was either 20% similar or 80% similar to the subject. Similarity was defined as being on the same side of the neutral point and one step away from the subject for the attitude. If the subject had marked the neutral point, the target also marked the neutral point. Dissimilarity was defined as being on the opposite side of the neutral point and four steps away from the subject unless the subject responded at the neutral point, in which case the target responded with either extreme agreement or disagreement with the statement—the side was randomly determined each time it occurred. Assignment to the 20 and 80% similarity conditions was randomly determined for all 62 students who had filled out the attitude questionnaire.

Three versions of a dossier were prepared to manipulate the target's competency, and one was randomly assigned to each subject. Each dossier described the target by indicating his birth date, April 22, 1932, in Ohio; he was married with two children and a graduate of Ohio State in 1954. He served in Korea for 3 years and received an honorable discharge. He worked 10 years with a competitor company. In a previous interview he was described as neat, well-dressed, and seemed favorably disposed to joining the new company; he had also filled out a questionnaire at that time. Competency was manipulated by indicating that the low competent (LC) target's past performance demonstrated a low work initiative and that he was not very creative in working but was punctual, usually meeting his deadlines. He had graduated from college with a "C" average, however, he often needed supervision. The average competent target (AC) had graduated, but no information about his grades was presented. He had average initiative, was not unusually creative, seldom missed deadlines but sometimes needed supervision. The high competent target (HC) had graduated with honors, had high initiative, was creative in his job performance, always met his deadlines and seldom needed supervision.

The subjects were run in groups of 3 to 15, and after they received the target's dossier, job recommendation form and attitude questionnaire, they were told to assume that they had been asked to evaluate a potential vice president by the company's president and to recommend the candidate's salary, from \$15,000 to \$25,000, even if the subject did not recommend hiring the candidate.

The Job Recommendation Form had seven 6-point Likert-type evaluations. The subjects evaluated how competent they felt the candidate was for the job, how strongly they recommended him for the job, how well he could handle an unusual situation, their desire for a personal interview with the candidate before deciding, how intelligent he was, his knowledge of current events, and how much he would cooperate with the company goals. The subjects were then asked to indicate the recommended salary and to rate him on five 6-step semantic differential adjectives: reliability, credibility, qualifiedness, valubleness and honesty.

### RESULTS

Each of the relevant responses made by the subjects was entered into a  $3 \times 2$  factorial unweighted means analysis of variance. The cell frequencies ranged from 7 to 10 subjects per cell. Table 1 presents the resulting summary table for each dependent variable. The first evaluation was that of the subjects' perception of the target's competency. This was a check on the manipulation of the target's competency. The manipulation was successful since the subjects rated the HC target most competent and the LC target least competent. However, the 80% similar target was perceived as significantly more competent than the 20% similar target.

It was predicted that both the similarity and competency effects would be significant contributors in the recommendation of hiring the target. The similarity effect failed to reach statistical significance even though the data tended to suggest that the effect was present and in the predicted direction. Thus, this part of the prediction was not confirmed. The more competent the target, the stronger was the recommendation, thus confirming this part of the prediction. Both similarity and competency were predicted to contribute to the amount of salary to be offered to the candidate. Both effects were significant. The 20% similar target received an average of \$16,923 whereas the 80% similar target received an average of \$18,126. The LC target received \$15,655, the AC target \$17,519, and the HC target \$19,188. These results confirm the predictions concerning the salary offered.

TABLE 1  
SUMMARY TABLE FOR EACH DEPENDENT VARIABLE

Source	df	Decision factors					
		Competence		Recommendation		Salary	
		MS	F	MS	F	MS	F
Similarity (A)	1	3.47	4.62**	3.63	3.15*	18.22	4.33**
Competency (B)	2	24.70	32.90***	29.74	25.80***	58.83	13.98***
A × B	2	0.54	0.72	1.89	1.64	3.27	0.78
Error	45	0.75		1.15		4.21	

\*  $p < .10$ .

\*\*  $p < .05$ .

\*\*\*  $p < .001$ .

## DISCUSSION

The first hypothesis that a more similar target would receive a stronger recommendation was not confirmed in the present experiment even though there was a tendency for the data to be in the predicted direction. All the candidates tended to receive rather low recommendations for the job opening and salaries, and these generally lower evaluations may have weakened the similarity effect. However, the less similar targets received a substantially lower salary than the high similar targets as predicted. Thus the general hypothesis has some support in the present data as in the Griffitt and Jackson (1970) study, for even though the disagreeing candidate may obtain the job, he is likely to receive a lower salary than the agreeing candidate. The lower salary may well prevent the person from accepting the job, which would have the same effect as not even offering him the job at all. In the event that a disagreeing candidate did accept the job at a lower salary, the inequity in pay might lead to greater absenteeism, dissatisfaction, disruption, or quitting—in each case, reinforcing the interviewer's belief that he should not have been hired.

The second hypothesis, that the more competent the target, the more likely he will be offered a job and at a higher salary was clearly confirmed in the present study. It is interesting to note, however, that the similar target was perceived as more competent than the dissimilar target. Exactly what is the role of attitude similarity in competency is not clear, but it seems to be related to the same processes that make a similar person seem more intelligent and knowledgeable about current events (Byrne, 1969).

The data provide additional information about the outcome of the interview process. Presum-

ably from the industrial point of view, the most desirable candidate would be the most competent. However, if the interviewer happens to disagree to a large extent with the candidate, he may be responsible for hiring a less competent person by controlling the salary offer. Clearly, this possibility must be considered as a potential aspect of the results of the interview process.

## REFERENCES

- BOLSTER, B. I., & SPRINGBETT, B. M. The reaction of interviewers to favorable and unfavorable information. *Journal of Applied Psychology*, 1961, 45, 97-103.
- BYRNE, D. Interpersonal attraction and attitude similarity. *Journal of Abnormal and Social Psychology*, 1961, 62, 713-715.
- BYRNE, D. Attitudes and attraction. In L. Berkowitz (Ed.), *Advances in experimental social psychology*. Vol. 4. New York: Academic Press, 1969.
- BYRNE, D., BASKETT, G. D., & HODGES, L. Behavioral indicators of interpersonal attraction. *Journal of Applied Social Psychology*, 1971, 1, 137-149.
- BYRNE, D., & CLORE, G. L. A reinforcement model of evaluative responses. *Personality: An International Journal*, 1970, 1, 103-128.
- CARLSON, R. E. Selection interview decisions: The relative influence of appearance and factual written information on an interviewer's final rating. *Journal of Applied Psychology*, 1967, 51, 461-468.
- GRIFFITT, W., & JACKSON, T. Influence of information about ability and non-ability on personnel selection decisions. *Psychological Reports*, 1970, 27, 959-962.
- MAYFIELD, E. C., & CARLSON, R. E. Selection interview decisions: First results from a long-term research project. *Personnel Psychology*, 1966, 18, 41-55.
- MILLER, J. W., & ROWE, P. M. Influence of favorable and unfavorable information upon assessment decisions. *Journal of Applied Psychology*, 1967, 51, 432-435.
- ROTTER, J. B. A new scale for the measurement of interpersonal trust. *Journal of Personality*, 1967, 35, 651-665.
- SPRINGBETT, B. M. Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology*, 1958, 12, 13-22.
- SYDIAHA, D. Interviewer consistency in the use of empathetic models in personnel selection. *Journal of Applied Psychology*, 1962, 46, 344-349.

(Received September 9, 1971)

## EFFECTS OF SPONSOR AND PREPAYMENT ON COMPLIANCE WITH A MAILED REQUEST<sup>1</sup>

ANTHONY N. DOOB<sup>2</sup>

*University of Toronto*

JONATHAN L. FREEDMAN

*Columbia University*

J. MERRILL CARLSMITH

*Stanford University*

A request to answer two questions on a stamped addressed postcard was sent to 804 people selected from a telephone directory. The sponsor of this request was either a university or a commercial firm. One third of the subjects received 20¢ with the request, one third received 5¢, and one third received no money. People were more likely to comply to the request from the university than the commercial firm, and compliance varied directly with the amount of money enclosed with the request. The difference between sponsors disappeared with increasing amounts of money.

A common problem in many areas of applied work is to get people to comply to a relatively small request. Clearly, one way of doing this is to apply coercive pressure to the person asked to comply. Happily or unhappily, this is not always possible, and other methods have to be devised. Freedman and Fraser (1966) found that when a person had been asked to comply to a small request before being asked to comply to a larger request, he was more likely to comply than if he had never been asked to comply to the small request. A person is also more likely to comply to any request if he has been made to feel guilty for something he has just done (Carlsmith & Gross, 1969; Freedman, Wallington, & Bless, 1967). Under certain conditions, compliance can be increased by having the person making the request appear stigmatized (Doob & Ecker, 1970).

In certain contexts, monetary prepayments appear to work to increase compliance. The theoretical reasons for this do not make much sense in terms of standard use of incentives, but may be understood in terms of the above mentioned findings on guilt. Kephart and Bressler (1958) found that increasing the amount of money (from no money to 25¢) enclosed with a request to fill out a questionnaire increased compliance from 52% to 70%. Similarly, Watson (1965) found increased compliance on a mail questionnaire from people who received money with the request. In addition, Doob and Zabrack (1971) found that

with various instructions, money included with a questionnaire tended to increase returns of the questionnaire. Apparently, then, including money with a questionnaire can increase compliance.

It is a fairly good assumption that the status of the requester also has an effect. Certainly, it is well documented that attitude change increases with the status of the communicator. In the area of compliance, status effects are not well documented. Moreover, there are no data on the combination of these two variables.

The present experiment combines the two variables—money payments in advance of compliance and sponsorship—into a three-by-two factorial design, with three levels of prepayment (0, 5¢, 20¢) and two sponsors of this request (Stanford University and Industrial Research Associates).

### METHOD

Sixty-seven pages from the San Mateo County, California telephone directory were chosen at random, and then 12 names were chosen at random from each of these pages with the restriction that they be the names of individuals and not businesses. Two names from each page were then assigned at random to each of the six conditions. Addresses were handwritten in ink. Enclosed was a mimeographed letter stating that "This is part of a survey which is being used to compare this area of the country with other areas. We would appreciate your filling out this questionnaire and returning it as soon as possible." The questionnaire consisted of a stamped preaddressed postcard with two questions ("Do you own an automobile?" and "Approximately how much television do you watch each day?") on it.

### Prepayment Manipulation

For subjects who were given 5¢ or 20¢ along with the request, an additional sentence was added: "We have included 5¢ (or 20¢) as a token of our thanks."

<sup>1</sup>The research reported here was supported by National Science Foundation grants to the second and third authors.

<sup>2</sup>Requests for reprints should be sent to Anthony N. Doob, Department of Psychology, University of Toronto, Toronto, Ontario, Canada M5S 1A1.

## Sponsors

Half of the subjects received requests from the Survey Research Center, Stanford University; the other half received requests from Industrial Research Associates with a post office box in an adjoining town.

## RESULTS AND DISCUSSION

A subject was considered to have complied if he answered the two questions and returned the postcard within two weeks. Results are based on  $n$ 's ranging from 125 to 133 per cell because of occasional letters that were returned by the post office because they could not be delivered. The results are shown in Figure 1. These proportions were transformed using the arcsin transformation and subjected to an analysis of variance. Both main effects and the interaction between the two variables were significant. Thus, people generally complied more to the University sponsor than to the commercial one ( $F = 6.84, p < .01$ ); people were more likely overall to comply when money was included ( $F = 7.35, p < .01$ ); and increasing amounts of money decreased the difference between the two sponsors (Interaction  $F = 5.35, p < .05$ ).

It is clear, then, that increasing the amount of money included with a request to fill out a questionnaire increases the rate of compliance to the request. Although this can be referred to as an "incentive," technically it is not, since the subject receives the money whether or not he fills out the questionnaire. The most plausible explanation for the increase in compliance with an increase in prepayment is that the subjects who receive money feel guilty if they accept the money and don't fill out the questionnaire. The subject, then, is placed in a bind where he will feel guilty unless he either complies or returns the money in his own envelope. In the Carlsmith and Gross (1969) and the Freedman, Wallington, and Bless (1967) experiments, guilt led to increased levels of compliance. In this experiment, increased levels of compliance were obtained because of an avoidance of guilt.

The practical application of these results is clear: Including a small amount of money with questionnaires will considerably increase compliance to a mailed request. This is especially true of a commercial sponsor for whom returns are normally relatively small.

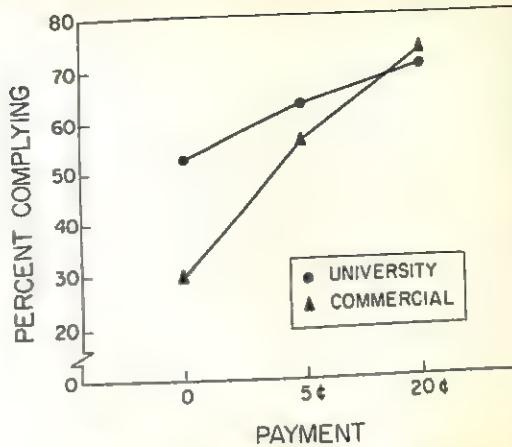


FIG. 1. Compliance as a function of sponsor and payment.

In our experiments he could increase his returns by over 90% (from 29.2% to 55.8%) simply by enclosing a nickel. Obviously, the increase in compliance with the addition of money will vary with a number of other characteristics (such as the sponsor, the size of the request, the subject population). The optimal amount of money to include to minimize the cost per response would have to be determined for each case.

## REFERENCES

- CARLSMITH, J. M., & GROSS, A. E. Some effects of guilt on compliance. *Journal of Personality and Social Psychology*, 1969, 11, 232-239.
- DOOB, A. N., & ECKER, B. T. Stigma and compliance. *Journal of Personality and Social Psychology*, 1970, 14, 302-304.
- DOOB, A. N., & ZABRACK, M. The effect of freedom-threatening instructions and monetary inducement on compliance. *Canadian Journal of Behavioural Science*, 1971, 3, 408-412.
- FREEDMAN, J. L., & FRASER, S. C. Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 1966, 4, 195-202.
- FREEDMAN, J. L., WALLINGTON, S. A., & BLESS, E. Compliance without pressure: The effect of guilt. *Journal of Personality and Social Psychology*, 1967, 7, 117-124.
- KEPHART, W., & BRESSLER, M. Increasing the response to mail questionnaires: A research study. *Public Opinion Quarterly*, 1958, 22, 123-132.
- WATSON, J. J. Improving the response rate in mail research. *Journal of Advertising Research*, 1965, 5, 48-50.

(Received November 24, 1971)

## EFFECTS OF SIGNED AND UNSIGNED QUESTIONNAIRES FOR BOTH SENSITIVE AND NONSENSITIVE ITEMS<sup>1</sup>

RICHARD P. BUTLER<sup>2</sup>

*United States Military Academy*

This study assessed the effects of signing and not signing questionnaires on items that were rated as not sensitive and on other items that were rated as sensitive. The subjects were 668 college seniors who responded to a mailed questionnaire covering a variety of different areas. Chi-square tests showed that there were no significant differences for any of the items between the respondents who signed and those who did not sign their questionnaires. It was concluded that the responses, despite variation in item sensitivity, were not influenced by signing or not signing questionnaires.

One of the most widely used methods to collect large amounts of data is by the use of questionnaires. When questionnaires are used, a problem frequently arises in regard to whether or not the respondents should be asked to identify themselves by signing their names. A number of investigators have stated that questionnaires should be administered anonymously if valid results are to be obtained (Corsini, 1948; Henle, 1949; Raphael, 1947; Stedman, 1947; Tiffin, 1950). However, it is often desirable to request the respondents to sign their names for purposes of making additional observations or data matching. If this is done, it is possible that identified respondents may distort their answers in the fear of some type of adverse effect. Typically, it is thought that when the information requested is of a sensitive nature, more distortion will occur than if the information sought is of a nonsensitive nature (Fischer, 1946).

The present study was an attempt to clarify the effects of signed and unsigned questionnaires, when the information requested was considered to be sensitive and when it was considered not sensitive. It was hypothesized that individuals who signed their questionnaires would have different response patterns than would individuals who remained anonymous for sensitive items, but that the response patterns would be similar for nonsensitive items.

<sup>1</sup> Any conclusions in this report are not to be construed as official United States Military Academy or Department of the Army positions unless so designated by other authorized documents.

<sup>2</sup> Requests for reprints should be sent to Richard P. Butler, Office of Institutional Research, United States Military Academy, West Point, New York 10996.

### METHOD

The questionnaires were sent to the 36 cadet company commanders at the United States Military Academy, who were responsible for distributing them to the seniors, collecting them when completed, and forwarding them to the office in charge of testing. The data for this study were collected by means of a questionnaire designed to gather information in a number of different areas; for example, attitudes toward drugs, racial problems, expectations of future Army life, curriculum, and demographic information. The respondents were given approximately 3 weeks to complete their questionnaires and to return them to their company commander. One third of the respondents were instructed to return their questionnaires unsigned, and the other two thirds were asked to sign their questionnaires. Of the 732 seniors involved, 668 (91%) responded.

To assess the degree of sensitivity of the data requested, three Department of the Army civilian psychologists, one United States Army officer, and one enlisted man (a clinical psychologist), all of whom were familiar with cadet life and practices, were asked to rate each of the items as to its degree of sensitivity from the viewpoint of the cadets. A 5-point scale was used with 0 equal to "not sensitive," 1 = "slightly sensitive," 2 = "definitely sensitive," 3 = "very sensitive," and 4 = "critically sensitive." The degree of sensitivity was defined as the extent to which an adverse effect (e.g., peer pressure, embarrassment, guilt, possible punitive or administrative action, etc.) on the cadet is possible in regard to how he answers the questions.

### RESULTS

To assess rater reliability, the intraclass correlation was employed on 30 randomly selected items. The extent of agreement among the five raters resulted in an intraclass coefficient of .31, which, when raised by the Spearman-Brown formula, becomes .69. This finding indicates that a fairly high degree of consistency among the five raters was obtained.

The majority of the items were given a low sensitivity rating. Twenty-eight of the

items received a sum of ratings score of 0, meaning that these items were rated by the judges as not sensitive. Seven items,<sup>3</sup> with summed rating scores of 9 through 12, received the highest ratings on sensitivity. Since the items in these two groupings obviously represented the extremes of sensitivity, they were chosen for closer analysis. Chi-square tests were made between the patterns of responses for the signed versus unsigned questionnaire for each of the seven items rated as most sensitive. No significant differences occurred at the .05 level of significance. The range of chi-square values was from 0.26 to 5.33, with *df* varying from 2 to 6.

For the sake of comparability, 7 items<sup>4</sup> were randomly chosen from the 28 that were judged as not sensitive. Chi-square tests were completed on these 7 items, and once again no significant differences were found at the .05 level for signed versus unsigned questionnaires. The chi-squares ranged from 1.43 to 5.94, with *df* varying from 1 to 4.<sup>5</sup>

### DISCUSSION

Contrary to what was hypothesized, the responses of respondents who signed their names and those who remained anonymous did not differ significantly for the items rated as most sensitive. However, as was expected, the responses of the two groups on items rated as not sensitive were also not significantly

<sup>3</sup> Of the seven items judged as most sensitive, two dealt with drugs, two with the cause of racial problems, one with minority cadet activity, one with career intention, and one with the wisdom of the cadets' original choice of attending the Military Academy.

<sup>4</sup> The seven least sensitive items that were chosen randomly were concerned with spring leave, special career programs, number of hours spent completing questionnaires, number of times asked to complete questionnaires by the academic departments, agreement with a hypothetical military academy degree in Military Management and Administration, expected satisfaction in the use of technical skills in their Army careers, and a Department of Army stipulation about call to active duty after resignation after the start of the third year.

<sup>5</sup> Tables presenting the frequency patterns of responses to both sensitive and nonsensitive items are available from the author on request (see address in Footnote 2).

different. Apparently, the degree of item sensitivity had little effect on the response patterns of the two groups.

A number of factors may have influenced the above results. One might speculate that the seven items judged as most sensitive, although rated as more sensitive than the other items, may not have really been sensitive in absolute terms. This explanation may have some credibility because the mean scores for these items were toward the middle of the rating scale. It might have been preferable to have items rated toward the highly sensitive end of the scale.

Another factor that may have been operating in this study was the fact that the cover letter to all respondents stated that no individual record of their responses would be made, that only group data would be analyzed, and that this information would be kept in the strictest confidence. This may have offset any fear that the respondents had in regard to an adverse effect if he signed his questionnaire. If this is so, then the instructions used in this study may be relevant for future studies that employ signed questionnaires.

In view of the results of this study, it appears that the use or nonuse of signatures is the option of the investigator, since no differences were noted in the answers of respondents who signed and those who did not sign their questionnaires. Of course, the degree of generality of these findings is a matter for empirical investigation.

### REFERENCES

- CORSINI, R. J. The pin prick method of secret balloting. *Journal of Applied Psychology*, 1948, 32, 641.
- FISCHER, R. P. Signed versus unsigned personal questionnaires. *Journal of Applied Psychology*, 1946, 30, 220-225.
- HENLE, D. R. Employee attitude survey: An analysis. *Personnel Journal*, 1949, 28, 218-225.
- RAPHAEL, W. The value of an attitude survey. *Journal of the Institute of Personnel Management*, 1947, 29, 275-280.
- STEDMAN, G. E. Industry audits. *Personnel Journal*, 1947, 25, 259-266.
- TIFFIN, J. The uses and potentialities of attitude surveys in industrial relations. *Proceedings of the Second Annual Meeting of the IRRA*, 1950, 2, 204-211.

(Received December 6, 1971)

## AUDITORY VIGILANCE UNDER HYPOXIA<sup>1</sup>

RICHARD L. CAHOON<sup>2</sup>

*United States Army Research Institute of Environmental Medicine, Natick, Massachusetts*

Vigilance performance by 12 young male subjects on a 2-hour loudness discrimination task was tested under four levels of hypoxia: 21% oxygen (sea level); 12.8% oxygen (13,000 feet); 11.8% oxygen (15,000 feet); and 10.9% oxygen (17,000 feet). The results indicated a significant decrement in signal detection as a function of severity of hypoxia and task duration. These findings paralleled those of visual vigilance studies and suggested that the function being affected by hypoxia is a central attention process rather than an orienting response.

Previous research in this laboratory has shown a decrement in visual vigilance performance at oxygen levels of less than 12.8%—13,000 feet altitude (Cahoon, 1970a, 1970b). Below this oxygen level, there was a significant decrease in the percentage of signals detected. There was also a significant decrease in  $d'$ , a measure from Signal Detection Theory (Tanner & Swets, 1954) that has been taken as an index of the sensitivity of a sensory system to the presence of a signal (Broadbent & Gregory, 1963).

Closed-circuit television observation of the behavior of subjects in these studies suggested that the process being affected by the hypoxia may not have been the ability to distinguish critical from noncritical signals, but rather the orienting behavior of the subjects. Reaction times for detected critical signals were as fast under conditions of hypoxia as they were at sea level. Time spent with the eyes open and fixated on the display, however, decreased considerably under hypoxia and could have accounted for the decrement found in percentage of signals detected. Jerison and Pickett (1963) have suggested that these orienting responses are special factors associated with visual monitoring and are made up of gross head and eye movements and keeping the eyes open. If it is these orienting responses that are disrupted by hypoxia, there should be less of a vigilance decrement in an auditory monitoring task. However, if in Jerison's terms, it is the "neural attention systems" that are affected by hypoxia, auditory and visual vigilance performance should be similar since these "systems" are common to both. The present study

was an attempt to distinguish between these possibilities.

### METHOD

#### Subjects

Twelve U.S. Army enlisted volunteers served as test subjects. Four oxygen levels (21% O<sub>2</sub>, 12.8% O<sub>2</sub>, 11.8% O<sub>2</sub>, and 10.9% O<sub>2</sub>) were administered to the subjects in a latin-square design with each subject undergoing all conditions. Total task duration was 2 hours during which data were compiled for each 30-minute segment.

#### Apparatus

A Massey-Dickenson behavioral programming system was used to present a 1,000 hertz tone to subjects on an AR speaker. The same equipment recorded and printed out the subject's response. The tone was set at 63 decibels of loudness for noncritical stimuli (nonsignals) and 67 decibels for critical stimuli (signals).

Hypoxia was induced by a means previously described (Cahoon, 1970a). The study was conducted in a soundproof chamber, and each subject was continuously monitored visually by way of a closed-circuit television system and auditorily by way of an intercom system.

#### Procedure

The subject sat facing the speaker that was set two feet above floor level. Tones were presented at the rate of  $\frac{1}{2}$  second on, 5 seconds off. The total duration of each trial was 2 hours during which each subject was given 36 signals (loud tones), 9 during each 30-minute period. All other stimuli were non-signals (soft tones). The subject indicated his detection of a signal by pressing a button that activated a printer. If the button was not pressed, the printer was activated automatically after 3 seconds.

Each subject ran a total of four trials, spaced 1 week apart, once at 21% O<sub>2</sub>, once at 12.8% O<sub>2</sub>, once at 11.8% O<sub>2</sub>, and once at 10.9% O<sub>2</sub>. Before each trial, a 5-minute practice trial was given in which the experimenter manually presented the subject with numerous signals and nonsignals. After each practice session, it was explained to subjects that the test trial

<sup>1</sup> Presented at the International Congress of Applied Psychology, Liege, Belgium, July 1971.

<sup>2</sup> Requests for reprints should be sent to Richard L. Cahoon, Behavioral Sciences Laboratory, U.S. Army Research Institute of Environmental Medicine, Natick, Massachusetts 01760.

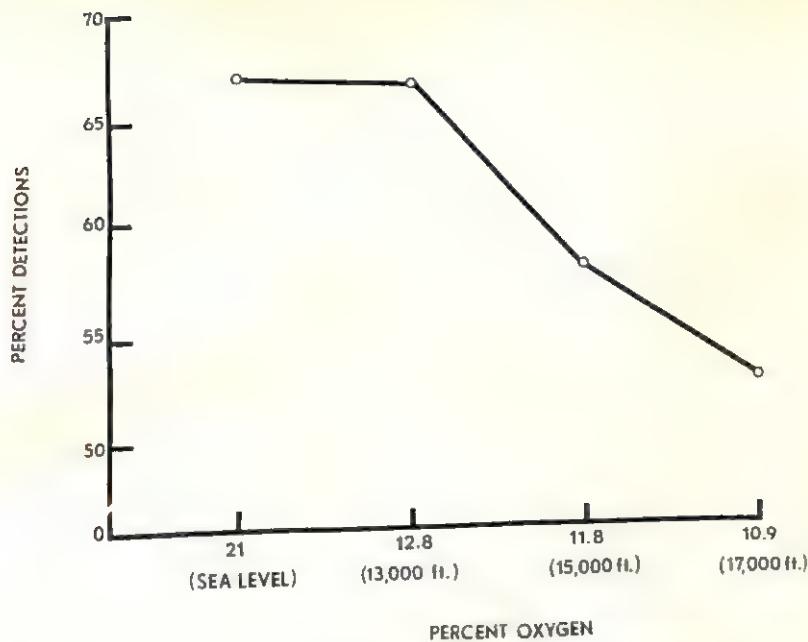


FIG. 1. Percent signals detected as a function of oxygen level.

schedule of signals was completely random and that each tone had an equal chance of being a signal or nonsignal. He was further told that it was possible that no signals at all would occur during the 2 hour trial. He was urged not to try to predict any schedule of signals but to listen to each tone and decide whether it was loud or soft. To guard against learning the pattern of signals, the schedule for presenting them was changed for each 30-minute period.

#### RESULTS

Data were compiled for each 30-minute segment and for the total 2 hour period. Measures taken of vigilance performance included percent signals detected, number of false detections, mean reaction time to signals, and mean reaction time to nonsignals (false detections). Each of these measures was analyzed by an ABS analysis of variance for the effects of hypoxia, and task duration. All subjects were represented in each analysis of variance, and there were no empty cells.

As in the visual vigilance studies, the primary measure of vigilance performance was the percentage of signals detected. The results of an analysis of variance of this measure showed a significant effect of hypoxia ( $p < .01$ ) with a sharp decrease occurring beyond the 13,000 foot level (see Figure 1). There was also a significant decrease ( $p < .01$ ) in percentage of signals detected as a function of task duration. There was a common finding in vigilance studies. There was

no significant interaction between hypoxia level and task duration.

Neither hypoxia level nor task duration had a significant effect on the number of false detections, the reaction time for correct detections, or the reaction time for false detections.

The measure,  $d'$ , was computed and found to decrease at a near significant rate ( $p < .10$ ) as a function of hypoxia level (see Figure 2). The measure,  $\beta$ , was also computed but showed no change with hypoxia.

#### DISCUSSION

A major purpose of the study was to determine whether presenting stimuli auditorily would improve performance over that found in visual vigilance studies. This did not occur even though orienting responses were not required to hear the critical signals. The auditory vigilance performance, although somewhat lower at sea level did, in fact, parallel the visual vigilance performance when subjects were hypoxic. Thus, the decrement in vigilance performance under hypoxia must be due to something other than interference with orienting responses.

The  $d'$  measure showed a decrease under hypoxia in the present study as in the visual studies, indicating again a possible decrease in the ability to detect the critical signals. However, as in the visual studies, the reaction times for detected

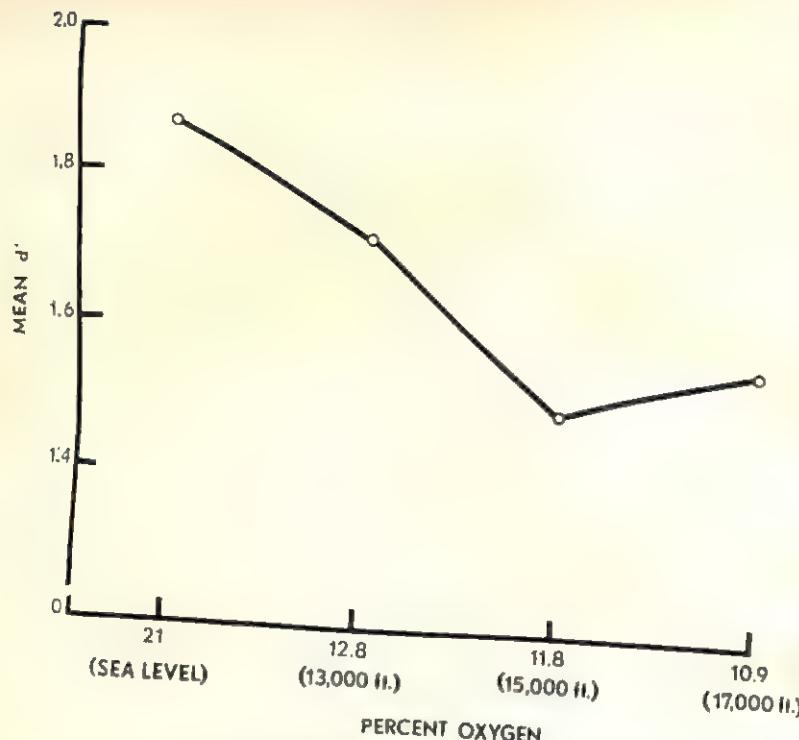


FIG. 2. Mean  $d'$  as a function of oxygen level.

signals did not differ between sea level and the various levels of hypoxia. Thus, the decrement in vigilance performance under hypoxia is probably due to something other than an inability to detect the difference in the intensity of the stimuli.

The data to this point suggest that the function being affected by hypoxia is an attention process that is central in nature and which corresponds to Jerison's (1963) "central neural attention mechanism." Although subjects in the present study need not have oriented to the display to detect a critical signal, they still had to attend to the stimulus. It appears that a major effect of the hypoxia was to decrease the ability of the subject to continuously attend to the stimulus for extended periods of time, possibly by attenuating the intensity of the neural messages that correspond to the signals or by lowering the general arousal level. Studies are currently in progress in this laboratory using cortical evoked potentials to determine the changes in response to sensory stimuli at the cortex that result from exposure

to hypoxia, both while the subject is attending and while he is not attending to the stimulus. It is hoped that these studies will shed further light on the properties of this basic attention process.

#### REFERENCES

- BROADBENT, D. E., & GREGORY, M. Vigilance considered as a statistical decision. *British Journal of Psychology*, 1963, 54, 309-323.
- CAHOON, R. L. Vigilance performance under hypoxia. *Journal of Applied Psychology*, 1970, 54, 479-483. (a)
- CAHOON, R. L. Vigilance performance under hypoxia: II. Effect of work-rest schedule. *Perceptual and Motor Skills*, 1970, 31, 619-626. (b)
- JERISON, H. J., & PICKETT, R. M. Vigilance: A review and re-evaluation. *Human Factors*, 1963, 5, 211-238.
- TANNER, W. P., & SWETS, J. A. A decision-making theory of visual detection. *Psychological Review*, 1954, 61, 401-409.

(Received October 4, 1971)

## PROMPTED MENTAL PRACTICE AS A FLIGHT SIMULATOR<sup>1</sup>

DIRK C. PRATHER<sup>2</sup>

*United States Air Force Academy*

Twenty-three subjects were randomly placed in one of two groups. All subjects were student pilots and minimally experienced in the landing of the T-37 aircraft, the independent variable. The experimental group (E) listened to four 12½-minute tape recordings that prompted their mental practice of landing the T-37 aircraft. The control group (C) did not receive this practice. All subjects were rated by their instructor pilots on procedures and ability to land the aircraft on the mission that followed the last mental practice session. Group E's ratings on both procedures and ability to land were significantly higher ( $p < .05$ ) than the ratings of group C. It was concluded that the use of mental practice may be an effective adjunct to any training program that normally depends on costly actual practice of the skill being learned.

With the rising cost of simulation devices, it is important to evaluate other devices and techniques that may be able to improve performance in a perceptual motor skill. Mental practice of a skill exists when the subject attempts to imagine vividly the perceptual motor actions involved in practicing the skill. Davis and Wallis (1961) have found that regular mental practice is superior to irregular actual practice in motor skill learning. Twining (1949) found no significant differences between actual and mental practice on basketball foul shooting. Shick (1970) was able to improve a volleyball skill through mental practice. Blurton (1969) used behavior therapy with imagery to improve significantly field goal shooting in practice, but found no significant differences in actual game situations. It appears that mental imagery can, in many cases, improve performance of a perceptual motor skill.

The author, in an unpublished study, attempted to improve strafing in student fighter pilots through mental practice. He found that mental practice of this skill did improve actual strafing scores over those that did not use the mental practice technique. Due to loss of control over the experimental subjects, statistical analysis was impossible.

Corbin (1967) found that some previous experience with the skill is necessary for mental practice to be effective. He decided that

landing an aircraft by low-experienced student pilots would be a skill in which the subjects had minimal experience, although a highly complex perceptual motor skill of the type that would be important to investigate. If this skill could be improved by mental practice, then it would strongly suggest that many less complex human skills may also be improved by this technique. This experiment was pointed toward improving performance in flight training in the United States Air Force. The subjects had some experience in landing an aircraft, but very little in landing the particular aircraft that was the independent variable. Due to the problems encountered in the control of the subjects in his pilot study, it was decided to use tape recordings as a prompt to the mental practice. This allowed for an exact timing of the student mental practice and a more precise control of his mental imagery. By weighing the student time and the cost of the apparatus, the cost effectiveness of such a program could be compared to more sophisticated methods of simulation.

The question proposed in this research was whether four highly prompted mental practice sessions of approximately 12½ minutes each could improve the student pilot's performance on landing an aircraft.

### METHOD

#### *Subjects*

The subjects were 23 randomly selected student pilots in the undergraduate T-37 pilot training program at Williams Air Force Base. Thirteen were in the experimental group, and 10 were randomly placed in the control group. All subjects were low-experienced student pilots with approximately 20 hours in the T-41 trainer and 4 hours in the T-37.

<sup>1</sup> The views expressed herein are those of the author and do not necessarily express the views of the United States Air Force or the Department of Defense.

<sup>2</sup> Requests for reprints should be sent to Dirk C. Prather, Department of Life and Behavioral Sciences, Department of the Air Force, USAF Academy, Colorado 80840.

### Apparatus

The experimental sessions for the experimental subjects were conducted in the learning center at Williams AFB. This center has typical student learning carrels for individual instruction through media presentation. The experimental subjects sat in a cockpit procedures trainer of the T-37 aircraft. This cockpit mock-up was configured similar to the actual aircraft through photographs. The only movable items in this mock-up were the throttles and the control stick. The instructions and stimulus information were played through earphones over a dial access tape recording.

### Procedure

The experimental subjects had observed and attempted the experimental task, that of landing the T-37 aircraft; but this experience was at a low level consisting of approximately seven previous landings. The experimental subjects were instructed to go to the learning center after they had completed the fourth, fifth, sixth, and seventh mission in the flying training syllabus and to listen to a tape recording while sitting in the cockpit mock-up.

The tapes were designed to give instruction in the landing pattern. The experimental subjects were told to imagine the situations as vividly as possible and to perform the same motor actions and eye movements that they would if they were in the actual landing pattern. In the first few imagined landing sequences the experimental subjects were given complete instructions as to the airspeeds, throttle settings, pitch attitudes, bank required, etc. In the later imagined patterns, the cues were withdrawn until in the last few sequences the tapes merely stated "You are on base" or "You are on final." To vary the sequences slightly, error analysis, go-arounds, touch-and-go, and final full-stop landings were all covered in this experimental training. The running time for each tape, in order, was 11:50, 15:10, 11:20, and 10:45.

The control subjects were not given any of the above experimental training. These control subjects received the normal training that past student pilots have received, which included some media presentations in the learning center.

After the eighth actual flying mission, both the experimental and control subjects were rated by their own instructor pilots on their performance as to technique and procedures in the landing pattern on that particular mission. This was a relative rating of the student's performance on several areas in the landing pattern. The instructor pilots did not know which students were in which group. Several instructor pilots had a student in each group to rate.

### RESULTS

The subjects' instructor pilots filled out a 1-7 rating scale on techniques and procedures

for the following phases of the landing: initial to pitch, pitch to 180, 180 to 10, final to flare, flare to touchdown, and go-around. The ratings for these phases of the landing pattern were averaged for each of the techniques and procedures area to give a more meaningful, stable rating. The procedures area was defined as how well the student knew what to do and the techniques area was defined as how well he actually did the landing task. The rating was relative in that the instructor was told to rate the subject in relation to all the other students he had instructed on that particular mission.

The results were analyzed by means of the Mann-Whitney *U* test. On procedures, the experimental group had a mean rating of 4.53 and the control group 4.26 ( $U = 35.3$ ,  $p < .05$ , two-tailed). On techniques, the experimental group had a mean rating of 4.21 and the control group 3.89 ( $U = 38.0$ ,  $p < .05$ , two-tailed).

### DISCUSSION

From the results of this experiment, it appears that mental practice combined with actual practice is more effective than just actual practice when learning a perceptual motor skill. The tape recorded presentation, using withdrawal of prompts to help control the mental imagery, is probably more effective than just letting the student imagine the skill without structure. Further structure was added to the mental practice by having the subject sit in the cockpit mock-up of the aircraft he was flying. With the extra practice gained by using prompts, it might be expected that the mental practice would improve the procedures of the subject; but the finding that the actual performance was improved through transfer of the skill practiced in the mental imagery sessions is very significant.

All experimental subjects filled out a critique on the program. Without exception they felt the mental practice helped them to perform better while flying. Most of the experimental subjects stated that they did not have any problem in vividly imagining the situations called for by the tape recordings.

Because the independent variable involved in this experiment is a highly complex perceptual motor skill, the results can probably be extended to include many areas of skill learning. The use of mental practice may be an effective, low-cost adjunct to any training program that

normally depends on costly actual practice of the skill being learned.

## REFERENCES

- BLURTON, R. R. Effects of group behavior therapy imagery on basketball performance. *Dissertation Abstracts*, 1969, 29, 3476-3477.
- CORBIN, C. B. The effects of covert rehearsal on the development of a complex motor skill. *Journal of Genetic Psychology*, 1967, 76, 143-150.
- DAVIS, E. C., & WALLIS, E. L. *Toward better teaching in physical education*. Englewood Cliffs, N. J.: Prentice-Hall, 1961.
- SNICK, J. Effects of mental practice on selected volleyball skills for college women. *Research Quarterly*, 1970, 41, 88-94.
- TWINING, W. E. Mental practice and physical practice in the learning of a motor skill. *Research Quarterly*, 1949, 20, 432-435.

(Received November 22, 1971)

*Journal of Applied Psychology*  
1973, Vol. 57, No. 3, 355-357

## EFFECTS OF PARTICIPATION IN A SIMULATED SOCIETY ON ATTITUDES OF BUSINESS STUDENTS

BENSON ROSEN,<sup>1</sup> THOMAS H. JERDEE, AND W. HARVEY HEGARTY<sup>2</sup>

*Graduate School of Business Administration, University of North Carolina*

An experiment was conducted to assess the effects of participation in SIMSOC (a simulated society involving social, economic, and political factors) on business students' attitudes toward business. It was found that SIMSOC participants placed greater emphasis on societal goals and less emphasis on suboptimizing business practices than did the control group.

In a recent Roper poll over half of the male college seniors throughout the United States thought business was "far too often not honest with the public," that it was "losing sight of values in the interest of profits," and that it was "hoodwinking the public through advertising" (Roper, 1969). This critical attitude of students toward business has caused business educators to become increasingly concerned about their role in shaping managerial attitudes.

The issues involved have been discussed by Schein (1967), who studied the attitudes of business school faculty members, graduate business students, and business executives. He found rather striking differences in attitude between faculty and executive groups. The attitudes of graduate students were initially somewhere between those of the faculty and executive groups on many issues, but during their period of college residence, the graduate students shifted away from the executive attitudes toward the faculty attitudes.

<sup>1</sup> Requests for reprints should be sent to Benson Rosen, Graduate School of Business Administration, University of North Carolina, Chapel Hill, North Carolina 27514.

<sup>2</sup> Now at the School of Business, West Virginia University, Morgantown, West Virginia.

Schein raised some interesting questions about this shift in attitude, including the permanence of the change, the difficulty it might cause students upon their entry into the business world, the business faculty's role as an attitude-change agent, and the efficacy of various attitude-change techniques.

Some degree of attitude change is a likely by-product of almost any educational experience. It certainly should occur in the educational game called SIMSOC, which is a simulation of society involving an array of business and social parameters. SIMSOC (simulated society) is a non-computerized simulation developed by Gamson (1969) which places students in various managerial roles where they are faced with conflicts between personal goals, organizational goals, and societal goals.

Participants are assigned various roles in the game: Some are business executives, union leaders, political party organizers, or members of the news media and the judicial council. Others are owners of travel agencies or subsistence agencies. Others have no ownership or leadership positions at the start of the game.

Organizational heads receive an income that they may use either for investment or to pay

people whom they hire as employees. The return on investments depends on the levels of national indicators. These indicators are a function of employment, investment in social welfare, and other variables.

Perhaps the outstanding features of SIMSOC, at least when the total society's resource base is set at a low level (as it was in this experiment), are the initially subtle, but eventually disastrous, consequences of individual and organizational disregard for the general condition of the whole SIMSOC society. In SIMSOC a strong emphasis on suboptimization results in misallocation of resources, poverty, unemployment, and ultimately in societal collapse. A game like SIMSOC may be viewed as a multiple role-playing situation, with a relatively high degree of realism. Therefore, it seems reasonable to assume that SIMSOC would influence attitudes in a manner similar to standard role-playing techniques (see e.g., Elms, 1967).

The salient features of SIMSOC which may be expected to have an impact on student attitudes are (a) the requirement that most participants enact attitude-discrepant roles, which might be expected to have the effect of attitude change in the direction of the position advocated, and (b) the informational aspects of the simulated depiction of the interrelationships between social and economic variables. It was therefore hypothesized that participation in SIMSOC would lead to attitudes of greater social awareness and concern.

#### METHOD

##### *Subjects*

The study was designed to compare SIMSOC participants with a control group. The subjects were 129 junior and senior undergraduate business students enrolled in four sections of a personnel problems course at the University of North Carolina.<sup>3</sup>

##### *Procedure*

Two sections of the course ( $n=42$ ) were combined for the purpose of participation in SIMSOC, while the two remaining sections acted as a control group. The simulation was enacted during six regular class meetings. During this time the control group attended lectures and class discussions. One week after the completion of SIMSOC, attitude questionnaires were administered to experimental and control groups. Since the SIMSOC game and the attitude questionnaire were treated as routine parts of the course, the students in the experimental group did not know that they were in an experiment and were not aware of the connection between the questionnaire and SIMSOC.

Attitudes were measured by a questionnaire con-

<sup>3</sup> The authors would like to thank D. J. Moffie for his cooperation in collecting the data.

taining 42 items, in the form of words or short phrases covering a wide variety of topics related to business. The items were arranged in six lists of seven, and the students were asked to rank the items in each list in terms of both ideal importance and actual importance to business. The two scores that were recorded for each person on each item were the ideal ranking and the ideal versus actual discrepancy. A subset of items were combined to form the following two scales: <sup>4</sup> (a) suboptimization—emphasis on individual self-enhancement and managerial achievement (items such as achievement, ambition, prestige, profit maximization, and growth of the business) and (b) concern for society—direct concern for the welfare of society as a whole (items such as cooperation, welfare of employees, welfare of society, tolerance, and justice).

Rankings of such "goals of business organizations" as efficiency, growth of the business, industry leadership, profit maximization, survival of the business, welfare of the business, welfare of employees, and welfare of society were included as separate measures of attitudes because of their special relevance in terms of the research hypothesis.

#### RESULTS

Two mean scores were derived for each scale and attitude item: (a) ideal—based on the students' rankings of the importance these items "ought to have" in society and (b) discrepancy—based on the difference between the ideal importance rankings and rankings of perceived actual importance to business managers. Although differences in "ideal" importance rankings were not significant when compared to the control group, SIMSOC participants assigned slightly higher ranks to items in the concern for society scale and lower ranks to items in the suboptimization scale.<sup>5</sup> In the discrepancy scores both groups perceived business managers as tending to under-emphasize concern for society and overemphasize suboptimization. The SIMSOC participants, however, perceived significantly greater discrepancies on both scales ( $t = 1.76, p < .05$ ;  $t = 2.27, p < .05$ ; respectively). Thus, it appears that SIMSOC tends to make students more critical of business, in the sense that they come to see a larger gap between their ideals and current business operative values.

In their rankings of the seven "goals of business organizations," SIMSOC participants as-

<sup>4</sup> These two scales were empirically derived on the basis of observed differences between American and Swedish business students. (See Jerdee, Brooks, & Barsk, 1971).

<sup>5</sup> Because of the ranking procedures, the  $t$  tests on the value measures are not completely independent. However, they are a useful means of identifying important differences.

signed significantly greater ideal importance to the welfare of society ( $t = 2.06, p < .05$ ). This finding provides direct support for our hypothesis that SIMSOC should broaden students' awareness of the consequences of business practices for society. They also saw a significantly greater overemphasis on growth ( $t = 2.40, p < .05$ ), and a significantly greater underemphasis on welfare of employees ( $t = 2.28, p < .05$ ). These findings add further support to our hypothesis.<sup>6</sup>

### DISCUSSION

Although the findings of the present study provide some support for the hypothesis that participation in SIMSOC leads to attitudes of greater social awareness and concern, it is difficult to assess the practical significance of these results since the differences are only marginally significant. The results, however, are encouraging for future work. Since many college students are already quite concerned with societal goals, attitude change for the student population may be limited by a "ceiling effect." This may be less likely to occur for business managers, since they generally tend to place less emphasis on societal goals.<sup>7</sup>

<sup>6</sup> Group means of "ideal" and "discrepancy" scores for the scales and individual attitude items are available upon request from the first author.

<sup>7</sup> A comparison with unpublished data collected by one of the authors on attitudes of 97 American business executives representing a geographically and industrially broad sample of companies indicates that executives score lower on concern for society and higher on suboptimization attitudes than control subjects in the present study. For further evidence see also Schein (1967), England (1967), and Roper (1969).

The effects of SIMSOC on students' activities outside the classroom should be explored. Rokeach (1971) has shown that classroom-induced shifts in values of freedom and equality are manifested in students' outside activities. Perhaps SIMSOC-induced shifts in managerial attitudes will be reflected in a similar fashion in later on-the-job behavior of participants. Follow-up studies would permit assessment of the nature, extent, and permanence of these effects, and the type of organizational climate most supportive of such new values.

### REFERENCES

- ELMS, A. C. Role-playing, incentive and dissonance. *Psychological Bulletin*, 1967, **68**, 132-148.
- ENGLAND, G. W. Personal value systems of American managers. *Academy of Management Journal*, 1967, **10**, 53-68.
- GAMSON, W. A. *SIMSOC: Simulated society*. New York: Free Press, 1969.
- JERDEE, T. H., BROOKS, W. W., & BARSK, B. Values among American and Swedish business students: Suboptimization vs. supraoptimization. *Acta Sociologica*, 1971, **14**, 256-267.
- ROKEACH, M. Long range experimental modification of values, attitudes, and behavior. *American Psychologist*, 1971, **26**, 453-459.
- ROPER RESEARCH ASSOCIATES INC. *The beliefs and attitudes of male college seniors, freshmen and alumnae*. New Jersey: Standard Oil Company, 1969.
- SCHEIN, E. H. Attitude change during management education. *Administrative Science Quarterly*, 1967, **11**, 601-628.

(Received November 17, 1971)

## NEUROTICISM AMONG POLICEMEN: AN EXAMINATION OF POLICE PERSONALITY

C. ABRAHAM FENSTER<sup>1</sup> AND BERNARD LOCKE

*John Jay College of Criminal Justice, City University of New York*

The neuroticism scores of 548 male subjects (college-educated policemen, non-college-educated policemen, and college- and non-college-educated civilians) were compared, using the Eysenck Personality Inventory and the Rokeach Dogmatism Scale. On the whole, policemen scored lower on neuroticism when compared with nonpolice citizens. Noncollege police were significantly less neurotic than noncollege civilians, and significantly less neurotic than college civilians on the Eysenck but not on the Rokeach scale. It was concluded that neuroticism was not a major characteristic of this group of policemen.

Previous research on police personality has indicated that emotional stability is a crucial factor in determining the probability of success of a policeman (Baehr, Saunders, Froemel, & Furcon, 1971). However, a considerable amount of research indicates that policemen may, in fact, be neurotic or even psychotic (Bain, 1939; Berman, 1971; Bohardt, 1959; Kates, 1950; Rankin, 1959; Rapaport, 1949; Skolnick, 1966; Westley, 1951; Zion, 1966).

The psychological literature is replete with suggestions that occupational choice represents, at least in part, a psychological defense against the recognition of certain unacceptable impulses in oneself (e.g., Roe, 1956; Shaffer & Shoben, 1956; Super, 1957). It has even been suggested that the occupational choice of becoming a policeman may be dictated by certain aggressive or authoritarian needs (Rapaport, 1949).

While it may well be that neurotic personalities do not perform well as policemen, most studies provide limited evidence about the actual level of neuroticism among policemen: (a) Many of the references implying neuroticism among policemen are speculations (albeit professional) that do not present psychometric or other research evidence. (b) While there is some evidence (Berman, 1971) that neurotics often apply for positions in the criminal justice field, very strict procedures for screening police applications exist in almost all major cities. (c) No direct comparisons are made between police neuroticism and that of civilians. (d) No distinction is made between the neuroticism scores of college-educated and noncollege-educated policemen. The purpose of the present study was to determine empirically

whether police and civilian groups at two educational levels differed in terms of neuroticism.

### METHOD

#### Subjects

A total of 548 male subjects were included in this study. The breakdown of this group is as follows: (a) 177 subjects were New York City patrolmen enrolled in various introductory psychology classes in a college of the City University of New York where about 50% of the student body were police officers attending college on a part-time basis. (b) 172 subjects were members of the New York City Police Department who worked with the first group and who were never enrolled in any college course. (c) 92 subjects were part-time students in introductory psychology classes at several other units of the City University of New York which had equivalent admission standards to the unit from which college-oriented police were chosen; these students were never policemen. (d) 107 subjects were adult civilians who had never been to college, and who had never served as policemen. The average ages of each group (noncollege police:  $\bar{X} = 32.17$ ,  $\sigma = 7.85$ ; college police:  $\bar{X} = 30.90$ ,  $\sigma = 7.45$ ; noncollege civilians:  $\bar{X} = 30.01$ ,  $\sigma = 10.20$ ; college civilians:  $\bar{X} = 25.55$ ,  $\sigma = 5.60$ ) were not significantly different except that college civilians were significantly younger than each of the other three groups ( $p < .01$ ).

#### Procedure

The Eysenck Personality Inventory (Form A) was administered to all subjects in several group sessions. Since it is often held that close-mindedness, rigidity, and dogmatism constitute the essence of neuroticism (or of psychopathology in general), the Rokeach Dogmatism Scale was administered to 545 of the 548 subjects as a secondary verification of the findings relating to neuroticism. (This scale has been used by other investigators as a measure of neuroticism—Kaplan & Singer, 1963; Yalom, 1970).

<sup>1</sup> Requests for reprints should be sent to C. Abraham Fenster, John Jay College of Criminal Justice, The City University of New York, 315 Park Avenue South, New York, New York 10010.

### RESULTS

The  $(4 \times 1)$  analysis of variance that was performed showed that Eysenck neuroticism scores among the four groups were significantly different ( $F = 9.16$ ,  $p < .001$ ). Several Duncan multiple-range tests, as adapted for unequal groups (Kramer, 1956), were performed in order to determine where these differences lay. The results were as follows:

1. The neuroticism scores of noncollege police ( $\bar{X} = 7.63$ ) are significantly lower ( $p < .001$ ) than those of noncollege civilians ( $\bar{X} = 10.47$ ).

2. The neuroticism scores of college-educated civilians ( $\bar{X} = 9.11$ ) are significantly lower ( $p < .05$ ) than those of noncollege-educated civilians ( $\bar{X} = 10.47$ ).

3. Noncollege civilians ( $\bar{X} = 10.47$ ) are significantly more neurotic than any of the other groups (college civilians,  $\bar{X} = 9.11$ ; college police,  $\bar{X} = 8.01$ ; noncollege police,  $\bar{X} = 7.63$ — $p < .001$  when compared with any police group and  $p < .05$  when compared with college civilians).

4. Noncollege police are significantly less neurotic than either college or noncollege civilians ( $p < .05$ ,  $p < .001$ ).

All results except the last hold true when the Rokeach Dogmatism scores are used as measures of neuroticism. In this instance there is no significant difference between noncollege police and college civilians.

### DISCUSSION

The results of this study clearly indicate that neuroticism is *not* a major characteristic of the average New York City policeman. On the whole, policemen scored lower than nonpolice citizens on the neuroticism scale of the Eysenck Personality Inventory and the Rokeach Dogmatism Scale. It is felt that other studies should be reevaluated in light of these tentative findings.

The applicability of these results was necessarily limited by the experimental design employed. Underwood (1957) has pointed out that whenever subject variables are being compared, randomization of all possible relevant variables is impossible. Only if subsequent research using different population samples confirms these results, can the original findings be made more tenable.

In light of Berman's (1971) recent study (which indicates that neurotics often apply for positions in the field of criminal justice), the present study

(which finds less neuroticism among police groups than is found in the general population) indicates that the intensive screening of police applicants, inadvertently or inadvertently, eliminates many neurotics. Because of the finding by Baehr et al. (1971) that emotional stability was a crucial factor in predicting good police performance, the authors believe that rigorous screening for neuroticism should be a part of police selection procedures.

### REFERENCES

- BAEHR, M. E., SAUNDERS, D. R., FROEMEL, E. C., & FURCON, J. E. The prediction of performance for black and for white police patrolmen. *Professional Psychology*, 1971, 2, 46-57.
- BAIN, R. The policeman on the beat. *Scientific Monthly*, 1939, 48, 452.
- BERMAN, A. MMPI characteristics of correctional officers. Paper presented at the meeting of the Eastern Psychological Association, New York, April 16, 1971.
- BOHARDT, P. H. Tucson uses new police personnel selection methods. *FBI Law Enforcement Bulletin*, 1959, 28, 8-12.
- KAPLAN, M. F., & SINGER, E. Dogmatism and sensory alienation: An empirical investigation. *Journal of Consulting Psychology*, 1963, 27, 486-491.
- KATES, S. L. Rorschach responses, strong blank scales and job satisfaction among policemen. *Journal of Applied Psychology*, 1950, 34, 252.
- KRAMER, C. Y. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, 1956, 12, 307-310.
- RANKIN, J. H. Psychiatric screening of police recruits. *Public Personnel Review*, 1959, 20, 191-196.
- RAPAPORT, D. *Diagnostic psychological testing*. Vol. 1. Chicago: Yearbook Publishers, 1949.
- ROE, A. *Psychology of occupations*. New York: Wiley, 1956.
- SHAFFER, L. F., & SHOBEN, E. J. *The psychology of adjustment*. Boston: Houghton Mifflin, 1956.
- SKOLNICK, J. H. *Justice without trial: Law enforcement in democratic society*. New York: Wiley, 1966.
- SUPER, D. *The psychology of careers*. New York: Harper, 1957.
- UNDERWOOD, B. J. *Psychological research*. New York: Appleton-Century-Crofts, 1957.
- WESTLEY, W. A. *The police: A sociological study of law, custom, and morality*. Unpublished doctoral dissertation, University of Chicago, 1951.
- YALOM, I. D. *The theory and practice of group psychotherapy*. New York: Basic Books, 1970.
- ZION, S. The police play a crime numbers game. *New York Times*, June 12, 1966, Section IV, 6.

(Received October 19, 1971)

## DOES FARM PRACTICE ADOPTION INVOLVE A GENERAL TRAIT?

JAMES M. RICHARDS, JR.<sup>1</sup>

University of Missouri-Kansas City

JOHN G. CLAUDY

American Institutes For Research, Palo Alto, California

Correlations among the adoption of recommended agricultural practices by agricultural workers were factor analyzed. The varimax solution indicated three identifiable factors that suggest a multidimensional nature of farm practice adoption. Implications for changing the behavior of agricultural workers with respect to specific practices are discussed.

In order to supply most of mankind with a minimally adequate diet, solutions must be found to the interrelated problems of controlling population growth and increasing food supply (Hutchinson, 1969). Psychologists are beginning to contribute to the finding of solutions to the problem of population growth (Pohlman, 1969) but continue to ignore almost entirely the problem of food supply. But, there is reason to believe (Brown, 1967; Nair, 1969) that increasing food supply is as much a problem in the applied psychology (i.e., ability, motivation, and knowledge) of those working in conventional agriculture as it is a problem in technology.

Achieving an understanding of the adoption of recommended farm practices appears, at least on the surface, especially amenable to attack from a psychological perspective. Specifically, considerable psychological methodology is available to study the assumption frequently made in studies of adoption that adoption behavior constitutes a general trait. But a review of *Psychological Abstracts* since 1930 revealed only two studies in the United States (Copp, 1956; Fliegel, 1956) explicitly testing this assumption (and those studies were by nonpsychologists). On the basis of moderate to high loadings for most included practices on the first unrotated principal component, these investigators concluded that there is a general trait of practice adoption.<sup>2</sup> From the perspective of more advanced computer technology and more recent factor-analytic methodology, it now appears that this conclusion may be oversimplified. Accordingly, the purpose of the present article is to reanalyze the

data from those two studies with a more "modern" factor-analysis procedure.

### METHOD

The study by Fliegel (1956) utilized data originally collected by Wilkening (1954) in 1952 from 170 farm owner-operators in Sauk County, Wisconsin. The Phi intercorrelation coefficients for adoption of the 11 farm practices shown in Table 1 were computed and the first principal component extracted. Since all loadings were greater than .2 and were fairly uniform in size, it was concluded that there was a general factor in evidence that could be called adoption of farm practices.

The study by Copp (1956) involved 157 cattle farmers in Wabaunsee County in the Flint Hills of Kansas. Again Phi intercorrelations for adoption of the 21 practices shown in Table 1 were computed. Because not all practices were applicable to all farmers, however, there was some variation in *Ns*. For each pair of variables, the correlation involved only farmers with a score on both variables. The first principal component was extracted from the intercorrelations among an eight-variable submatrix including the most highly intercorrelated practices that did not involve great expense and did not apply only to cowherds. Because the lowest loading was .48, and because the off-diagonal residuals were all low and negative, it was concluded that this principal component expressed a general predisposition to adopt recommended practices.

The present study involved refactoring of these intercorrelation matrices. In addition to a different method of factor analysis, the procedure departed from that followed in these studies in two important ways. First, to obtain some estimate of the effect of limitations on the size of Phi resulting from the marginal proportions, each correlation coefficient from the Fliegel study was transformed by the Phi/Phi Max. correction.<sup>3</sup> (Any overall bias of this transformation appears to be in the direction of generality). Second, all 21 practices from the Copp study were included in the analysis.

Each intercorrelation matrix was factored by the principal axes method. For each variable, the diagonal value was the squared multiple correlation between all other variables and that variable. In each case, the

<sup>1</sup> Requests for reprints should be sent to James M. Richards, Jr., Office of Medical Education, University of Missouri at Kansas City, Kansas City, Missouri, 64108.

<sup>2</sup> In a personal communication (1971), Copp reports obtaining similar results in studies in India.

<sup>3</sup> The authors thank Frederick C. Fliegel for supplying information about the proportion of farmers adopting each practice.

"Scree" test (Cattell, 1966) for discontinuities in the curve of eigenvalues indicated that three factors should be retained in the final solution. Accordingly, the intercorrelation matrices were refactored by the principal axes method, using as diagonal values the communalities computed from the first three unrotated factors. Three factors were extracted from each intercorrelation matrix and rotated to final solutions by the varimax procedure.

## RESULTS AND DISCUSSION

The rotated matrices are shown in Table 1. These factors are briefly described and interpreted below:

For the Fliegel study, Factor A has its highest loading on use of artificial insemination and use of a registered sire. There is a sizeable drop to the next highest loading. The common element in these variables is obviously insemination; indeed there may even be a part-whole relationship between the two variables. An obvious title would be *Insemination Practices*. Factor B has its highest loadings on clipping udders, use of a milking machine, and use of a mechanical milk cooler. Because of the evident common element, *Milking Practices* would be a good title. Factor C has high loadings on use of residual fly spray, use of high nitrogen fertilizer as a side dressing, and recent extensive use of fertilizer. *Use of Chemicals* might be the most appropriate title.

For the Copp study, Factor A has high loadings on feeding minerals, keeping a feed reserve, lice and grub control, fly control, and use of terracing. All of these practices were among the 8 analyzed originally by Copp (1956) who stated that they have been recommended for a long time, are not prohibitively costly, and in general are clearly economically justifiable. The best title for this Factor, therefore, might be *Conservative Good Practices*. Factor B has high loadings on castrating calves, dehorning calves, use of Blackleg vaccination, use of Bang's vaccination, and pen bull. Nearly all of these practices pertain to calves, so *Care of Calves* would seem an appropriate title. Factor C has high loadings on pen bull, having a soil test, use of fertilizer, fly control, trial of the deferred system of beef production, lice and grub control, and water close. An unambiguous interpretation is not immediately evident, for two alternatives appear plausible. First, it is possible these variables have been recommended more recently, so a good title would be *Receptivity to Change*. Alternatively, it is possible that these practices are less central to the main enterprise,

TABLE 1  
VARIMAX ROTATED FACTOR MATRICES

Variable	Factor			
	A	B	C	$h^2$
Fliegel study				
Use of fertilizer	.30	-.03	.38	.23
Recent soil test	.22	.11	.50	.31
High nitrogen fertilizer on corn	.42	.05	.62	.57
Use of registered sire	.69	.37	.14	.63
Clipping of udders	.12	.70	.02	.50
Use of haybaler	-.06	.06	.26	.08
Use of 2-4-D	.09	.11	.15	.04
Artificial insemination	.79	-.02	.03	.62
Milking machine	.07	.66	.15	.46
Mechanical milk cooler	-.05	.43	.34	.30
Use of fly spray	.01	.18	.68	.50
Copp study				
Purebred bull	.31	.01	.14	.12
Pen bull	-.02	.33	.53	.38
Creep feed calves	.13	.15	.29	.12
Castrate calves	-.08	.75	.05	.57
Dehorn calves	.01	.65	.07	.43
Blackleg vaccination	.07	.33	.06	.12
Bang's vaccination	.34	.46	.14	.34
Protein in ration	.06	.00	.33	.11
Minerals in ration	.58	.25	.05	.40
Spring pasture	.06	.10	.11	.03
Brush control	.16	.19	.00	.06
Water close	-.24	-.07	.36	.19
Ponds	.18	.08	.04	.04
Feed reserve	.45	-.02	.03	.20
Lice and grub control	.45	.07	.36	.34
Fly control	.40	.29	.37	.38
Recent soil test	.22	.09	.45	.26
Use of fertilizer	.31	.05	.45	.30
Legumes in rotation	.24	.03	.30	.15
Use of terraces	.40	-.01	.18	.19
Trial of deferred system of beef production	.32	.22	.37	.28

Note. Analyses are based on correlations among these variables obtained by Fliegel (1956) and Copp (1956).

and therefore something like *Peripheral Good Practices* would be the best title.

Together, these results cast strong doubts on the usefulness of treating farm practice adoption as a general trait. Any bias in the "Scree" test does not appear to be in the direction of too many factors, and naturally a three-factor solution accounts for more variance than a one-factor solution. Moreover,

the rotated factors are reasonably clear and interpretable. In the Fliegel study, these factors can be readily identified with three separate, broad areas of farm technology. The results do not, of course, demonstrate that Fliegel and Copp were wrong in any absolute sense in the conclusion that practice adoption is a general trait. In the field of mental abilities, g theories are a legitimate alternative to multitrait theories, and the first unrotated component does measure g. Nevertheless, the results do suggest that practice adoption can be treated more appropriately and usefully as complex and multidimensional than as general. (Additional studies aimed directly at the question of generality might be helpful). From a practical point of view, the results also suggest that efforts designed to change farmer behavior with respect to specific practices should in turn be specific. That is, rather than relying on, or seeking to inculcate, a generalized receptive attitude in farmers toward adoption of recommended practices, each practice should be "sold" to farmers on the basis of its own merits in helping them achieve their goals as farmers.

Such criterion complexity may also be characteristic of other aspects of farmer performance, and may explain, in part, studies (Richards, 1972) that fail to find much rela-

tionship between psychological measures and measures of success in farming.

#### REFERENCES

- BROWN, L. R. The world outlook for conventional agriculture. *Science*, 1967, 158, 604-611.
- CATTELL, R. B. The Scree test for the number of factors. *Multivariate Behavioral Research*, 1966, 1, 245-276.
- COPP, J. H. Personal and social factors associated with the adoption of recommended farm practices among cattlemen. (Tech. Bull. No. 83) Manhattan, Kan.: Agricultural Experiment State, Kansas State University, 1956.
- FLIEGEL, F. C. A multiple correlation analysis of factors associated with adoption of farm practices. *Rural Sociology*, 1956, 21, 284-292.
- HUTCHINSON, J. (Ed.) *Population and food supply*. Cambridge, England: Cambridge University Press, 1969.
- NAIR, K. *The lonely furrow: Farming in the United States, Japan, and India*. Ann Arbor: University of Michigan Press, 1969.
- POHLMAN, E. *The psychology of birth planning*. Cambridge, Mass.: Schenkman, 1969.
- RICHARDS, J. M., Jr. A longitudinal study in a national sample of young farmers in the United States. *Journal of Vocational Behavior*, 1972, 2, 179-189.
- WILKENING, E. A. Change in farm technology as related to familism, family decision making, and family interaction. *American Sociological Review*, 1954, 19, 29-37.

(Received December 20, 1971)

